مقاله پژوهشی

ترکیب تکنیک‌های انتخاب نمونه و داده‌افزایی برای حل مسئله طبقه‌بندی مجموعه داده‌های نامتوازن

پرستو محقق، سمیرا نوفرستی و مهری رجائی

چکیده: در عصر کلان‌داده‌ها، تکنیک‌های تجزیه و تحلیل خودکار مانند داده‌کاوی به‌طور گسترده‌ای برای تصمیم‌گیری به‌کار گرفته شده و بسیار مؤثر واقع شده‌اند. از جمله تکنیک‌های داده‌کاوی می‌توان به طبقه‌بندی اشاره کرد که یک روش رایج برای تصمیم‌گیری و پیش‌بینی است. الگوریتم‌های طبقه‌بندی به‌طور معمول بر روی مجموعه داده‌های متوازن به‌خوبی عمل می‌کنند. با وجود این، یکی از مشکلاتی که الگوریتم‌های طبقه‌بندی با آن مواجه هستند، پیش‌بینی صحیح برچسب نمونه‌های جدید بر اساس یادگیری بر روی مجموعه داده‌های نامتوازن است. در این نوع از مجموعه داده‌ها، توزیع ناهمگونی که داده‌ها در کلاس‌های مختلف دارند باعث نادیده گرفته‌شدن نمونه‌های کلاس با تعداد نمونه کمتر در یادگیری طبقه‌بند می‌شوند؛ در حالی که این کلاس در برخی مسائل پیش‌بینی دارای اهمیت بیشتری است. به‌منظور مقابله با مشکل مذکور در این مقاله، روشی کارا برای متعادل‌سازی مجموعه داده‌های نامتوازن ارائه می‌شود که با متعادل‌نمودن تعداد نمونه‌های کلاس‌های مختلف در مجموعه داده‌ای نامتوازن، پیش‌بینی صحیح برچسب کلاس نمونه‌های جدید توسط الگوریتم یادگیری ماشین را بهبود می‌بخشد. بر اساس ارزیابی‌های صورت‌گرفته، روش پیشنهادی بر اساس دو معیار رایج در ارزیابی طبقه‌بندی مجموعه داده‌های نامتوازن به نام‌های «صحت متعادل» و «ویژگی»، عملکرد بهتری در مقایسه با روش‌های دیگر دارد.

کلیدواژه: **انتخاب نمونه، داده‌افزایی، طبقه‌بندی، مجموعه داده نامتوازن، داده‌کاوی، یادگیری ماشین.**

۱- مقدمه

حجم کلان‌داده‌ها به‌گونه‌ای عظیم است که چالش‌های ذخیره‌سازی و تحلیل داده‌ها، کشف دانش و پیچیدگی محاسباتی را به دنبال دارد. از جمله کلان‌داده‌ها می‌توان به پروژه ژنوم انسان[1] اشاره کرد که چندین گیگابایت داده را از کد ژنتیکی انسان تجمیع کرده است. داده‌کاوی، مجموعه‌ای از روش‌های قابل اعمال بر مجموعه داده‌های بزرگ و پیچیده به‌منظور کشف الگوهای پنهان در میان داده‌هاست.

در میان تکنیک‌های مختلف داده‌کاوی، طبقه‌بندی[2] از تکنیک‌هایی است که بیشترین استفاده را برای مسائل مختلف داراست؛ مانند پیش‌بینی ورشکستگی و تشخیص سرطان [۱] و [۲]. طبقه‌بندی یکی از روش‌های یادگیری ماشین است که بر اساس یک مجموعه داده آموزشی از نمونه‌های برچسب‌خورده به ساخت یک مدل پیش‌بینی‌کننده می‌پردازد که قادر است برچسب کلاس نمونه‌های جدید را تعیین کند.

یکی از مشکلاتی که الگوریتم‌های طبقه‌بندی با آن مواجه هستند، مجموعه داده‌های نامتوازن است. مشکل عدم توازن زمانی رخ می‌دهد که تعداد نمونه‌ها در یک کلاس (که کلاس اقلیت[3] نامیده می‌شود) در مقایسه با سایر کلاس‌ها بسیار کمتر باشد؛ در حالی که کلاس اقلیت از لحاظ کاربرد اهمیت فراوانی دارد. به عنوان مثال می‌توان به مسائل زیر اشاره کرد. در شناسایی آفت‌های کشاورزی، برخی آفت‌ها پررخداد هستند و به‌کرات نمونه دارند و برخی آفت‌ها به‌ندرت مشاهده می‌شوند و به همین دلیل، تعداد نمونه‌های کمی دارند [۳]. در تشخیص نفوذ به شبکه‌های کامپیوتری (تشخیص ترافیک نرمال از ترافیک غیرنرمال شبکه) و تشخیص اشیا- جایی که اغلب تصاویر مجموعه آموزش فاقد شیء مورد نظر هستند- نیز برخی از کلاس‌ها تعداد نمونه‌های کمی دارند [۴].

برای حل مشکل مجموعه داده‌های نامتوازن، رویکردهای متعددی در سه سطح داده، مدل و ترکیبی ارائه شده‌اند. هدف رویکردهای سطح داده کاهش نسبت عدم تعادل بین کلاس‌های اکثریت[4] و اقلیت با انتخاب نمونه[5] کلاس اکثریت یا داده‌افزایی[6] در کلاس اقلیت است [۵] و [۶]. رویکردهای سطح مدل عمدتاً توابع حساس به هزینه[7] را معرفی می‌کنند [۷]. رویکردهای این سطح غالباً شامل روش فرایادگیری[8] و روش یادگیری انتقالی[9] هستند [۸] و [۹]. روش‌های ترکیبی نیز مزایای رویکردهای سطح داده و سطح مدل را ترکیب می‌کنند.

در این مقاله بر رویکرد سطح داده برای حل مشکل داده‌های نامتوازن متمرکز می‌شویم. رویکردهای سطح داده که به‌طور گسترده استفاده می‌شوند با متعادل‌سازی مجموعه داده نامتوازن، بهبود نتایج یادگیری ماشین را به دنبال دارند. الگوریتم‌های طبقه‌بندی سنتی بر روی داده‌های نامتوازن عملکرد مناسبی ندارند؛ زیرا فرض می‌کنند که تعداد نمونه‌های

این مقاله در تاریخ ۱۹ اسفند ماه ۱۴۰۱ دریافت و در تاریخ ۹ مرداد ماه ۱۴۰۲ بازنگری شد.

پرستو محقق، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران، (email: P.mohaghegh@pgs.usb.ac.ir).
سمیرا نوفرستی (نویسنده مسئول)، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران، (email: snoferesti@ece.usb.ac.ir).
مهری رجائی، دانشکده مهندسی برق و کامپیوتر، دانشگاه سیستان و بلوچستان، زاهدان، ایران، (email: rajayi@ece.usb.ac.ir).

1. Human Genome
2. Classification
3. Minority Class
4. Majority Class
5. Instance Selection
6. Data Augmentation
7. Cost-Sensitive Functions
8. Meta-Learning
9. Transfer Learning

آموزشی از هر کلاس با هم یکسان هستند و طبقه‌بند را بر مبنای این فرض آموزش می‌دهند. در واقع، زمانی که تعداد نمونه‌ها در کلاس اکثریت بیشتر از کلاس اقلیت باشد، الگوریتم یادگیری ماشین بیشتر به کلاس اکثریت توجه می‌کند و کلاس اقلیت را نادیده می‌گیرد؛ در حالی که در بسیاری از مسائل واقعی، پیش‌بینی صحیح برچسب نمونه‌های کلاس اقلیت دارای اهمیت بیشتری است. بنابراین این امر سبب پیش‌بینی ضعیف نمونه‌های کلاس اقلیت می‌شود؛ زیرا کلاس اقلیت به‌درستی آموزش داده نشده است. برای حل مشکل مذکور تکنیک‌های متعددی برای متعادل‌سازی مجموعه داده‌های نامتوازن معرفی شده‌اند.

به‌طور کلی، تکنیک‌های موجود 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 دارند [۱۰] و [۱۱]. در روش‌های انتخاب نمونه، گاهی بخش عمده‌ای از اطلاعات که برای تهیه و برچسب‌گذاری آنها زمان و هزینه زیادی صرف شده است، از دست می‌روند و مشکل انتخاب نمونه بیش از حد رخ می‌دهد که منجر به کاهش عملکرد طبقه‌بند می‌شود.

در روش‌های داده‌افزایی، الگوریتم‌های نمونه‌برداری به‌صورت تصادفی یا از مناطقی که اهمیت بیشتری برای کاربر دارد، نمونه‌های کلاس اقلیت را کپی‌برداری می‌کنند تا زمانی که تعادل بین نمونه‌های کلاس اقلیت و اکثریت حاصل شود [۱۲] و [۱۳]. در عین سادگی این روش‌ها، مشکلی که وجود دارد این است که نمونه مورد نظر از کلاس اقلیت برای انجام تکرار مشخص نیست.

برای اجتناب از مشکلات ذکرشده، در این مقاله روشی ترکیبی ارائه می‌شود که از مزایای هر دو روش انتخاب نمونه و داده‌افزایی بهره می‌برد. در روش پیشنهادی، ابتدا با ترکیب دو روش داده‌افزایی به گسترش مجموعه اقلیت پرداخته می‌شود و سپس با به‌کارگیری یک روش انتخاب نمونه مبتنی بر اتوماتای یادگیر[1]، نمونه‌هایی از کلاس اقلیت گسترش‌یافته گسترش‌یافته که تأثیر بیشتری بر دقت مدل پیش‌بینی‌کننده دارند، انتخاب می‌شوند. نتایج آزمایش‌های انجام‌گرفته، کارایی روش پیشنهادی را در مقایسه با روش‌های موجود نشان می‌دهند.

ادامه مقاله به‌صورت زیر سازمان‌دهی شده است: در بخش ۲ به معرفی تحقیقات پیشین در زمینه طبقه‌بندی داده‌های نامتوازن پرداخته می‌شود. در بخش ۳ جزئیات روش پیشنهادی شرح داده می‌شود. در بخش ۴ ارزیابی کارایی روش پیشنهادی و مقایسه نتایج آن با سایر روش‌های موجود ارائه می‌گردد و در پایان، بخش ۵ نتیجه‌گیری است.

۲- مرور تحقیقات پیشین

تکنیک‌های متعادل‌سازی مجموعه داده‌های نامتوازن در کارهای پیشین را به‌طور کلی می‌توان به دو دسته انتخاب نمونه و داده‌افزایی تقسیم کرد.

۲-۱ تکنیک‌های انتخاب نمونه

تکنیک‌های انتخاب نمونه به دو دسته کلی فیلتر و رپر دسته‌بندی می‌شوند. تکنیک‌های فیلتر اغلب به رویکرد نزدیک‌ترین همسایگان یک نمونه جهت حذف یا حفظ آن وابسته هستند. روش‌های فیلتر بر اساس تابع انتخاب نمونه عمل می‌کنند و از لحاظ محاسباتی بسیار سریع‌تر از روش‌های رپر هستند و برای مجموعه داده‌ها با ابعاد بزرگ نیز مناسب هستند. از جمله این روش‌ها می‌توان به روش پایه POP [۱۴] اشاره کرد. این روش، نمونه‌های مرکزی را حذف و نمونه‌های مرزی را حفظ می‌کند. در متد POP مقدار Weakness برای هر یک از نمونه‌ها حساب می‌شود که نشانگر تعداد دفعاتی می‌باشد که نمونه مد نظر به‌عنوان نمونه مرکزی شناخته شده است. پس از آن، نمونه‌هایی که مقدار Weakness برایشان برابر تعداد ویژگی‌ها باشد، به‌عنوان نمونه مرکزی شناخته می‌شوند و حذف می‌گردند. در [۱۰] یک روش فیلتر برای انتخاب نمونه پیشنهاد شده که بر بهینه‌کردن سه معیار همبستگی، افزونگی و سازگاری داده‌ها به‌صورت همزمان تمرکز می‌کند. در این روش، قیود حفظ نمونه بر مدل‌های بهینه‌سازی تحمیل می‌شوند تا حداکثر درصد نمونه‌هایی را که توسط تصمیم‌گیرنده ایجاد شده است حفظ کنند.

از جمله روش‌های رپر پایه می‌توان به مدل‌های DROP۱-۵ از ویلسون و مارتینز [۱۵] اشاره کرد که در آن، ترتیب حذف نمونه‌ها اهمیت دارد. بدین صورت که اگر حذف یک نمونه P از مجموعه کاهش‌یافته S بر دقت طبقه‌بند تأثیری نداشته باشد، آن نمونه از S حذف می‌شود. این مدل‌ها به‌ویژه DROP۳ پایه بسیاری از روش‌های جدید انتخاب نمونه هستند [۱۱].

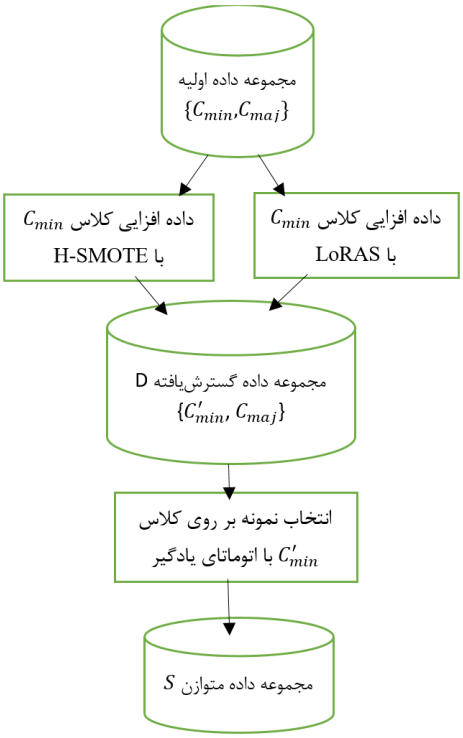
تیسای و همکاران [۵] روشی را با نام CBIS[2] پیشنهاد دادند که در آن، آن، مجموعه آموزش نامتوازن با دو کلاس اکثریت و اقلیت بررسی می‌گردد. در گام اول از کلاس اقلیت صرف نظر می‌شود و الگوریتم خوشه‌بندی، نمونه‌های مشابه از کلاس اکثریت را در تعدادی خوشه گروه‌بندی می‌کند که به‌عنوان زیرکلاس‌هایی از کلاس اکثریت شناخته می‌شوند. سپس در گام بعد، الگوریتم انتخاب نمونه پایه مانند الگوریتم ژنتیک بر روی کلیه خوشه‌ها عمل فیلتر را انجام می‌دهد و دو مجموعه داده کاهش‌یافته با نمونه‌های نویزدار و بدون نویز ایجاد می‌شوند. مجموعه با نمونه‌های نویزدار کنار گذاشته می‌شود و مجموعه داده کاهش‌یافته نهایی بدون نویز انتخاب می‌گردد. سپس مجموعه داده کاهش‌یافته نهایی از کلاس اکثریت با نمونه‌های کلاس اقلیت ترکیب شده و به‌عنوان مجموعه آموزشی جدید به طبقه‌بند جهت ارزیابی عملکرد داده می‌شود. نتایج تجربی نشان می‌دهند بعد از اعمال روش CBIS بر روی مجموعه داده نامتوازن، نسبت عدم توازن مجموعه تا حدی کاهش می‌یابد.

در [۱۶] روشی مبتنی بر یادگیری تقویتی برای انتخاب نمونه پیشنهاد شد که ابتدا خوشه‌هایی از نمونه‌ها را بر اساس یک معیار شباهت می‌سازد. سپس عاملی را در نظر می‌گیرد که می‌تواند تدریجاً داده‌ها را به مجموعه کاهش‌یافته نهایی اضافه کند. در یک حلقه، این عامل خوشه‌ای از نمونه‌ها (عمل) را برای اضافه‌شدن به نمونه‌های انتخاب‌شده (حالت)، انتخاب 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 و این ماتریس توانایی متعادل‌کردن اهداف کاهش نویز و حجم داده را دارد.

در مقاله حاضر نیز یک تکنیک یادگیری تقویتی به نام اتوماتای یادگیر برای انتخاب نمونه پیشنهاد شده و نتایج مطالعات پیشین، موفقیت‌آمیزبودن به‌کارگیری اتوماتای یادگیر را برای حل مسائل بهینه‌سازی پیچیده که شامل عدم قطعیت، غیرخطی‌بودن و متغیرهای تصمیم‌گیری چندگانه

1. Learning Automata

2. Cluster-Based Instance Selection



شکل ۱: مراحل روش پیشنهادی.

هستند، نشان می‌دهد. در زمینه انتخاب نمونه، اتوماتای یادگیر می‌تواند طی یک فرایند تکرارشونده به یافتن زیرمجموعه بهینه (یا نزدیک به بهینه) از نمونه‌ها با تنظیم مکرر آنها بر اساس بازخورد دریافتی از محیط کمک کند. از جمله مزایای اتوماتای یادگیر، عدم نیاز به دانش صریح درباره مسئله بهینه‌سازی است؛ مثل اطلاعاتی مانند ساختار مسئله و محدودیت‌هایی که باید برآورده شوند. علاوه بر این، اتوماتای یادگیر می‌تواند با کاوش مؤثرتر در فضای راه‌حل‌های ممکن، مسائل با ابعاد بالا را به نحو مؤثرتری مدیریت کند.

۲-۲ تکنیک‌های داده‌افزایی

ایده اصلی در روش‌های داده‌افزایی، گسترش تعداد نمونه‌های کلاس اقلیت برای متوازن‌سازی توزیع نمونه‌ها بین کلاس‌ها است. روش SMOTE[1] در بین روش‌های داده‌افزایی از محبوبیت بیشتری برخوردار است [۱۷]. ایده اصلی این روش به این صورت است که در کلاس اقلیت برای هر نمونه P با توجه به میزان شباهتی که با دیگر نمونه‌های کلاس اقلیت دارد، k نزدیک‌ترین همسایگانش از درون کلاس اقلیت تعیین می‌گردند و سپس بین هر دو نمونه از همسایگان، نمونه‌های جدید به‌صورت تصادفی تولید می‌شوند. مشکلی که روش SMOTE دارد این است که نمی‌تواند به‌خوبی بر مشکل توزیع نامتوازن داده‌ها غلبه کند؛ به دلیل اینکه ناحیه تولید نمونه‌ها (کلاس اقلیت) محدود است و فاصله نمونه‌های مرکزی تولیدشده در مجموعه داده‌ای جدید با نمونه‌های مرکزی مجموعه داده آموزشی اصلی زیاد می‌شود.

چائو و ژانگ [۱۲] نیز روش H-SMOTE را با هدف تولید نمونه‌های جدید یکنواخت‌تر و نزدیک‌تر به مرکز کلاس اقلیت معرفی کردند. در این روش عملیات صورت‌گرفته به‌طور عمده از دو عمل نمونه‌برداری و فیلتر نمونه‌های نویزدار 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 و نمونه جدید در سطح اول تولید می‌گردد. این عمل برای نمونه‌های جدید و نمونه‌های کلاس اقلیت در سطح دوم نیز تکرار می‌شود. گرچه روش H-SMOTE عملکرد بهتری نسبت به الگوریتم‌های پایه مانند SMOTE دارد، اما همچنان مشکل کافی‌نبودن نمونه‌های تولیدشده در سطح اول و دوم برای متوازن‌سازی با افزایش نرخ عدم توازن را دارد.

در تحقیقی دیگر، بیج و همکاران [۱۳] یک روش داده‌افزایی را با نام LoRAS[2] معرفی کردند. در این روش برای هر نمونه P از کلاس اقلیت، محلی چندضلعی در نظر گرفته شده و سپس k نزدیک‌ترین همسایه‌های نمونه P مطابق محل آن انتخاب می‌شوند. برای هر نمونه که در همسایگان نمونه P قرار دارد به اندازه معینی، نمونه سایه ایجاد می‌گردد. نمونه‌های سایه به‌وسیله تعدادی نویز با توجه به توزیع نرمال با انحراف معیار استاندارد ایجاد می‌شوند. سپس به‌صورت تصادفی به تعداد ویژگی‌ها از لیست نمونه‌های سایه، نمونه انتخاب می‌گردد و در وزن‌های بردار آفین ضرب می‌شوند و خروجی بردار به‌عنوان نمونه جدید در نظر گرفته می‌شود. اگر F یک میدان و V یک فضای برداری روی F باشد که اعضای F اسکالر و v_iها (که $i=1,2,\ldots,n$) اعضای فضای برداری V باشند، آن گاه ترکیب خطی $V=\lambda_1 v_1+\lambda_2 v_2+\ldots+\lambda_n v_n$ یک ترکیب آفین نامیده می‌شود اگر $\sum\lambda_j=1$ باشد [۱۸].

مسئله متعادل‌سازی مجموعه داده‌های نامتوازن، همچنان مسئله‌ای چالش‌برانگیز برای پژوهشگران است. در این مقاله یک روش ترکیبی مبتنی بر داده‌افزایی و انتخاب نمونه با درنظرگرفتن اهمیت حفظ نمونه‌ها بر مبنای عملکرد طبقه‌بندها ارائه شده است.

۳- روش پیشنهادی

روش پیشنهادی متمرکز بر بهره‌مندی از مزایای هر دو روش انتخاب نمونه و داده‌افزایی برای افزایش کارایی الگوریتم‌های یادگیری ماشین است. مطالعات کارهای پیشین نشان داده که انتخاب نمونه و داده‌افزایی، تأثیر شایانی در افزایش دقت الگوریتم‌های یادگیری ماشین دارد. با وجود این، استفاده از روش‌های انتخاب نمونه و داده‌افزایی به‌طور مستقل مشکلاتی را به همراه دارد. روش‌های انتخاب نمونه موجب حذف نمونه‌هایی می‌شوند که برای جمع‌آوری و برچسب‌گذاری آنها وقت و هزینه بسیاری صرف شده و روش‌های داده‌افزایی منجر به تولید نمونه‌های تکراری متعدد و غیرکاربردی می‌شوند. لذا برای حل مشکل ذکرشده در پژوهش حاضر، روش ترکیبی مبتنی بر هر دو روش انتخاب نمونه و داده‌افزایی برای متوازن‌ساختن دو کلاس اقلیت و اکثریت مجموعه داده و همین طور بهبود دقت الگوریتم‌های یادگیری ماشین ارائه می‌شود.

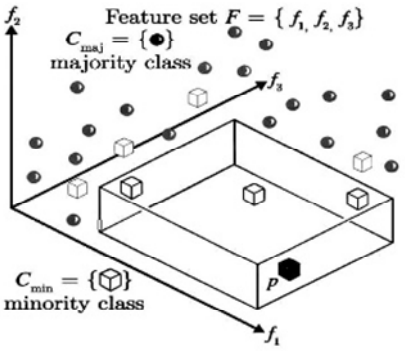
روش پیشنهادی برای متوازن‌سازی کلاس اقلیت و اکثریت که به‌ترتیب با $C_{\min}$ و C_{maj} نشان داده می‌شوند، دو مرحله اصلی دارد: ۱) داده‌افزایی کلاس اقلیت $C_{\min}$ با ترکیب دو روش شناخته‌شده H-SMOTE [۱۲] و LoRAS [۱۳] که منجر به ایجاد یک مجموعه جدید از رکوردهای آموزشی به نام کلاس اقلیت گسترش‌یافته $(C'_{\min})$ می‌شود و ۲) انتخاب نمونه با اتوماتاهای یادگیر بر روی کلاس $C'_{\min}$. مراحل روش پیشنهادی در شکل ۱ نشان داده شده‌اند. در ادامه جزئیات هر مرحله تشریح می‌گردد.

۳-۱ مرحله اول: داده‌افزایی

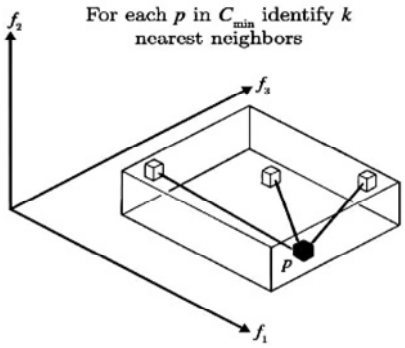
روش‌های داده‌افزایی، سعی در افزایش نمونه‌های کلاس اقلیت با هدف

1. Synthetic Minority Over Sampling Technique

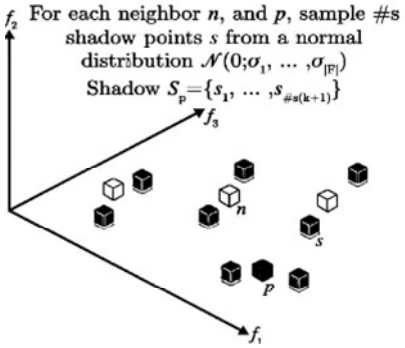
2. Localized Random Affine Shadow Sampling



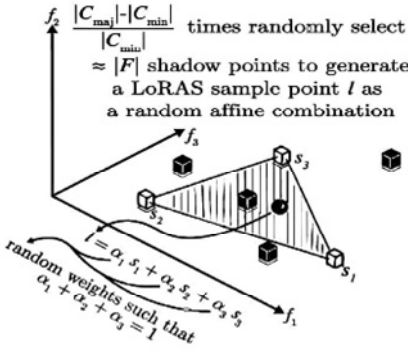
(الف)



(ب)



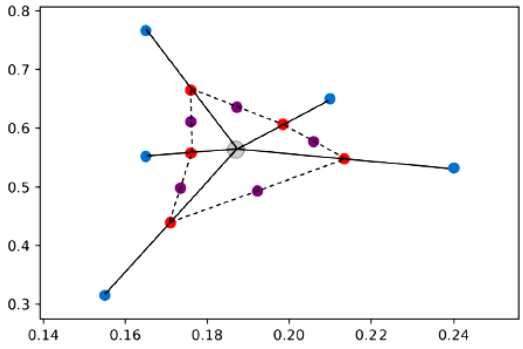
(ج)



(د)

شکل ۲: مصورسازی گام‌به‌گام روش داده‌افزایی LoRAS [۱۳].

متعادل‌سازی توزیع نمونه‌ها بین کلاس‌ها دارند. از آنجا که روش پیشنهادی مقاله حاضر، سعی در متعادل‌سازی کلاس اقلیت با کلاس اکثریت دارد، ابتدا داده‌افزایی کلاس اقلیت با دو الگوریتم LoRAS [۱۳] و H-SMOTE [۱۲] صورت می‌گیرد. دلیل انتخاب دو الگوریتم مذکور، محبوبیت و نتایج موفقیت‌آمیز آنها در مقایسه با روش‌های موجود است. الگوریتم H-SMOTE دو مزیت اصلی دارد: توزیع پایدار و یکنواخت نمونه‌های تولیدشده و کارایی الگوریتم در مقایسه با روش‌های موجود [۱۲]. الگوریتم LoRAS همچنین علاوه بر کارایی بالا، توانایی مدیریت



شکل ۳: داده‌افزایی مبتنی بر الگوریتم H-SMOTE [۱۲].

مجموعه داده‌های بسیار نامتوازن و با ابعاد بالا را دارد [۱۳]. در ادامه به معرفی دقیق‌تر این دو الگوریتم می‌پردازیم.

شکل ۲ الگوریتم داده‌افزایی LoRAS را نمایش می‌دهد. در این الگوریتم در فضای F بعدی (که F تعداد ویژگی‌هاست) برای هر نمونه p از کلاس اقلیت که با $C_{\min}$ نمایش داده می‌شود، محلی t ضلعی در نظر گرفته شده که به‌صورت پیش‌فرض، چندضلعی منتظم مانند مکعب مستطیل است. سپس k نزدیک‌ترین همسایه‌های نمونه p مطابق محل آن انتخاب می‌شوند. برای هر نمونه q والد که در همسایگان نمونه p قرار دارند، مطابق (۱) به اندازه $|S_p|$ نمونه سایه ایجاد می‌شود

$$|S_p| = \max\left(\left\lceil \frac{2|F|}{k} \right\rceil, 40\right) \quad (1)$$

که مقدار ۴۰ حداکثر تعداد نمونه‌های سایه برای یک نمونه والد را نشان می‌دهد. نمونه‌های سایه به وسیله تعدادی نویز با توجه به توزیع نرمال با انحراف معیار استاندارد L_σ ایجاد می‌شوند و مقدار پیش‌فرض آن $[0.005,\ldots,0.005]$ می‌باشد. سپس به‌صورت تصادفی به اندازه N_{aff} که پیش‌فرض تعداد ویژگی‌ها است از لیست نمونه‌های سایه، نمونه انتخاب شده و طبق (۲) در وزن‌های بردار آفین ضرب می‌شوند و خروجی بردار به‌عنوان نمونه جدید به مجموعه نمونه‌های LoRAS افزوده می‌شود [۱۳]. یک بردار آفین با وزن‌های تصادفی به‌صورت $\alpha_1 + \alpha_2 + \alpha_3 = 1$ تعریف می‌گردد و نمونه جدید از طریق (۲) به‌دست می‌آید

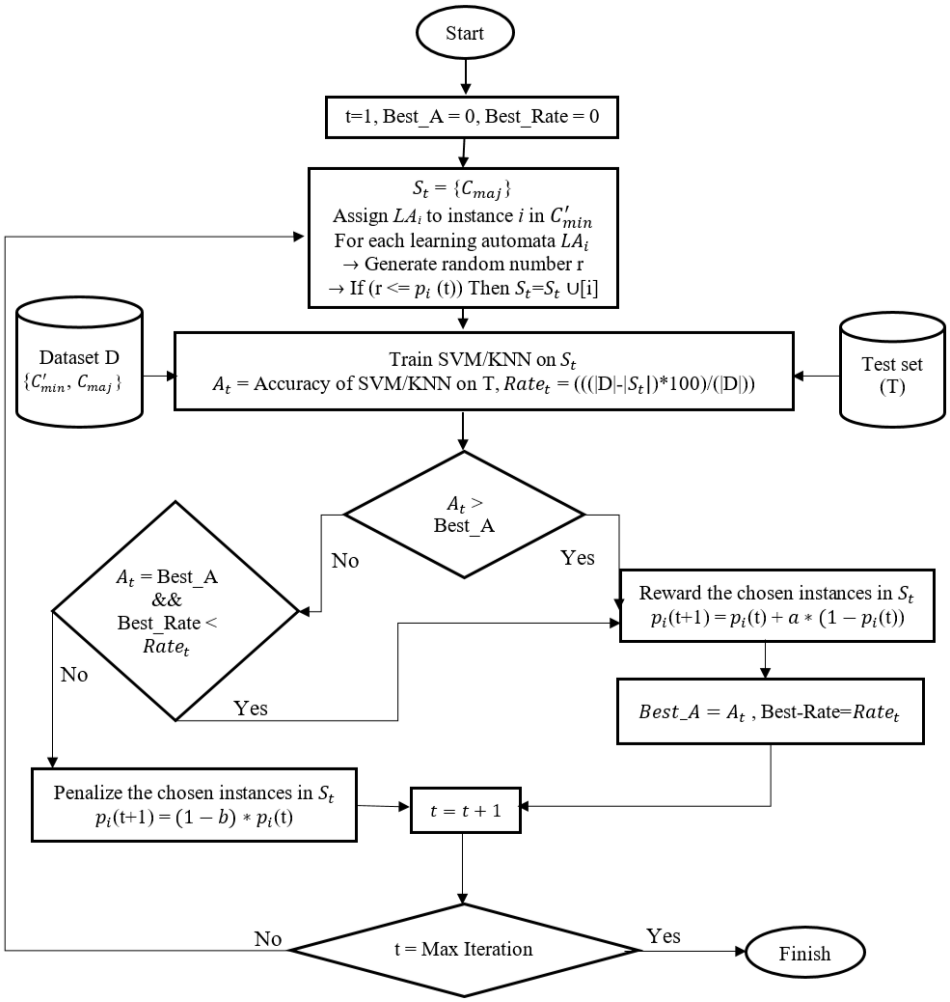
$$L = \alpha_1 s_1 + \alpha_2 s_2 + \alpha_3 s_3 \quad (2)$$

این الگوریتم تا زمانی که تعداد نمونه‌ها برای هر گروه نزدیک‌ترین همسایگی مطابق (۳) به اندازه N_{gen} نرسد، اجرا می‌شود که $|C_{\min}|$ تعداد نمونه‌های کلاس اقلیت و $|C_{maj}|$ تعداد نمونه‌های کلاس اکثریت را نشان می‌دهد

$$N_{gen} = \frac{|C_{maj}| - |C_{\min}|}{|C_{\min}|} \quad (3)$$

در الگوریتم H-SMOTE، نمونه‌های اقلیت و اکثریت به‌عنوان ورودی به الگوریتم داده می‌شوند و خروجی الگوریتم، مجموعه داده‌ای از جنس کلاس اقلیت است که نمونه تکراری در آن وجود ندارد. مطابق شکل ۳، در ابتدا برای به‌دست‌آوردن نمونه مرکزی Z در کلاس اقلیت، میانگین نمونه‌های کلاس اقلیت (نقاط آبی) محاسبه می‌شود. سپس در سطح اول برای هر نمونه p_i از کلاس اقلیت مطابق (۴) با محاسبه فاصله منهتن، نمونه جدید x_i میان نمونه مرکزی Z و نمونه p_i تولید می‌شود و در کلاس اقلیت قرار می‌گیرد (نقاط قرمز) [۱۲]

$$x_i = p_i + 0.5 \times (Z - p_i) \quad (4)$$



شکل ۴: روندنمای الگوریتم انتخاب نمونه.

از آنجا که تعداد نمونه‌های تولیدشده در سطح اول برابر با تعداد نمونه‌های کلاس اقلیت است، اگر تعداد نمونه‌های کلاس اکثریت و نمونه‌های کلاس اقلیت را به‌ترتیب A و B نام‌گذاری کنیم، آنگاه تعداد نمونه‌های تولیدشده در سطح دوم باید $A-۲\times B$ باشد. بنابراین به‌طور تصادفی دو نمونه x_i و x_j از نمونه‌های سطح اول که مساوی نیستند برای محاسبه فاصله منهتن انتخاب می‌شوند و سپس مطابق (۵) نمونه جدید x_{ii} در سطح دوم ایجاد می‌گردد (نقاط بنفش) [۱۲]

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

نهایتاً نمونه‌های تکراری، فیلتر و نمونه‌های جدید حاصل می‌شوند. این الگوریتم تا زمانی که تعداد نمونه‌های جدید، بزرگ‌تر یا مساوی کلاس اکثریت نباشد، ادامه می‌یابد تا همان طور که گفته شد، تعداد نمونه‌های سطح دوم برابر $A-۲\times B$ شود.

همان طور که گفته شد، داده‌افزایی تنها بر روی کلاس اقلیت $C_{\min}$ انجام می‌شود و حاصل آن کلاس اقلیت گسترش‌یافته $C'_{\min}$ است. از آنجا که هر دو روش LoRAS و H-SMOTE سعی دارند تعداد نمونه‌های کلاس اقلیت را به تعداد نمونه‌های کلاس اکثریت برساند، با اعمال توأم هر دو روش به تعادل نسبی در تعداد نمونه‌های کلاس‌های اقلیت و اکثریت نمی‌رسیم؛ بلکه تعداد نمونه‌های کلاس اقلیت گسترش‌یافته $C'_{\min}$ بیشتر از تعداد نمونه‌های کلاس اکثریت C_{maj} می‌شود. برای حل این مشکل، پس از داده‌افزایی با دو روش LoRAS و H-SMOTE، مجموعه داده خروجی آنها ترکیب شده (که آن را D می‌نامیم) و برای مرحله دوم (انتخاب نمونه) به‌کار گرفته می‌شود. مجموعه D شامل کلیه رکوردهای $C'_{\min}$ و C_{maj} بوده که فراوانی برچسب کلاس $C'_{\min}$ بیشتر است. به همین دلیل برای رسیدن به تعادل در اندازه کلاس‌ها در مرحله دوم سعی می‌شود با انتخاب زیرمجموعه‌ای از رکوردهای کلاس $C'_{\min}$ در مجموعه D که برای طبقه‌بندی مؤثرتر هستند و حذف رکوردهای نویز و زائد، در تعداد نمونه‌های دو کلاس اقلیت و اکثریت تعادل نسبی برقرار شود.

۳-۲ مرحله دوم: انتخاب نمونه

انتخاب نمونه یکی از مراحل پیش‌پردازش داده است که به دنبال انتخاب زیرمجموعه‌ای باکیفیت از کل داده‌های موجود با حذف رکوردهای نویز و زائد با هدف بهبود دقت الگوریتم‌های طبقه‌بندی است. در این مقاله برای انتخاب نمونه از اتوماتای یادگیر استفاده شده است. اتوماتای یادگیر مدلی است که به‌طور تصادفی یک عمل را از مجموعه متناهی از اعمال انتخاب و در محیط اِعمال می‌کند. سپس محیط عمل انتخاب‌شده توسط اتوماتای یادگیر را ارزیابی می‌نماید و نتیجه ارزیابی خود را توسط یک سیگنال تقویتی به اتوماتای یادگیر اطلاع می‌دهد. اتوماتای یادگیر با دریافت این سیگنال تقویتی، وضعیت خود را به‌روز کرده و عمل بعدی خود را انتخاب می‌کند [۱۹].

هدف استفاده از اتوماتای یادگیر، انتخاب زیرمجموعه‌ای باکیفیت از رکوردهای کلاس اقلیت گسترش‌یافته $C'_{\min}$ است. در واقع با حذف رکوردهای نویز و زائد از کلاس $C'_{\min}$ (که فراوانی بیشتری در مقایسه با کلاس C_{maj} دارد)، به دنبال ایجاد تعادل نسبی در اندازه کلاس اقلیت و اکثریت هستیم. در شکل ۴ روندنمای الگوریتم انتخاب نمونه آمده است.

فرایند کار بدین صورت است که به هر نمونه (رکورد) i در کلاس $C'_{\min}$ از مجموعه داده ترکیبی (D) حاصل از مرحله اول، یک اتوماتای یادگیر LA_i تعلق می‌گیرد. هر اتوماتای LA_i بر اساس احتمال $p_i(t)$ که متناظر با عمل انتخاب نمونه i در تکرار tام است، یک عمل (انتخاب یا عدم انتخاب نمونه متناظر) را برمی‌گزینند. برای این منظور در تکرار tام، اتوماتای یادگیر LA_i که متعلق به نمونه i در کلاس $C'_{\min}$ است، یک مقدار تصادفی r را تولید می‌کند و اگر $r < p_i(t)$ باشد، نمونه i در مجموعه کاهش‌یافته S_t قرار می‌گیرد که این مجموعه، حاصل الگوریتم در پایان تکرار tام است. در ابتدا هر نمونه از کلاس $C'_{\min}$، احتمالی یکسان (برابر ۰/۵) برای انتخاب یا عدم انتخاب جهت افزوده‌شدن به مجموعه کاهش‌یافته S_t دارد.

در پایان هر تکرار t، یک مجموعه کاهش‌یافته به نام S_t حاصل می‌شود که شامل کلیه نمونه‌های کلاس C_{maj} و نمونه‌هایی از کلاس $C'_{\min}$ است که توسط اتوماتاهای یادگیر در تکرار t انتخاب شده‌اند. یک الگوریتم طبقه‌بندی بر روی مجموعه کاهش‌یافته S_t آموزش می‌بیند و به پیش‌بینی برچسب مجموعه داده تست T که بر اساس روش ارزیابی متقابل $10-\text{fold}$ به‌دست آمده است، می‌پردازد. کارایی طبقه‌بند بر اساس معیار صحت[1] بر روی مجموعه S_t محاسبه می‌شود و همین طور نرخ کاهش نیز حساب شده است و بر اساس نتایج به‌دست‌آمده به نمونه‌های منتخب پاداش یا جریمه تعلق می‌گیرد. بدین صورت که اگر مقدار صحت طبقه‌بند بر روی مجموعه کاهش‌یافته فعلی از بیشترین مقدار به‌دست‌آمده برای معیار صحت در تکرارهای قبلی $(Best\_A)$ بیشتر باشد، بردار احتمال‌های اتوماتاها به‌روزرسانی شده و عمل انتخابی آنها پاداش دریافت می‌کند. اگر مقدار صحت طبقه‌بند با $Best\_A$ یکسان باشد و نرخ کاهش در این تکرار از نرخ کاهش در تکرارهای قبلی $(Best\_Rate)$ بیشتر باشد نیز عمل انتخابی اتوماتاهای یادگیر پاداش می‌گیرد. اگر خلاف این موارد باشد، آن گاه عمل انتخابی اتوماتاهای یادگیر جریمه می‌شود. بنابراین احتمال انتخاب نمونه‌های مجموعه کاهش‌یافته S_t که کارایی طبقه‌بند مد نظر را بر اساس معیار صحت بالا برده‌اند، در تکرارهای بعدی افزایش یافته و احتمال انتخاب نمونه‌های مجموعه کاهش‌یافته S_t که باعث افت کارایی طبقه‌بند بر اساس معیار صحت شده‌اند در تکرارهای بعدی کاهش می‌یابد. پاداش و جریمه اتوماتاهای یادگیر به‌ترتیب مطابق (۶) و (۷) انجام می‌شود

$$\begin{aligned} p_i(n+1) &= p_i(n) + a[1 - p_i(n)] \\ p_j(n+1) &= (1-a)p_j(n) \quad \forall j,\ j \neq i \end{aligned} \tag{۶}$$

$$\begin{aligned} p_i(n+1) &= (1-b)p_i(n) \\ p_j(n+1) &= \frac{b}{r-1} + (1-b)p_j(n) \quad \forall j,\ j \neq i \end{aligned} \tag{۷}$$

که a و b دو مقدار ثابت هستند که به ترتیب پارامتر پاداش و جریمه نامیده می‌شوند. در این مقاله با سعی و خطا مقادیر این دو پارامتر ۰/۰۸ و ۰/۰۰۵ در نظر گرفته شده است. تا زمانی که شرط توقف (رسیدن به حداکثر تکرارها) اجرایی شود، فرایند فوق ادامه خواهد داشت. در پایان الگوریتم، بهترین راه حل حاصل به‌عنوان مجموعه کاهش‌یافته نهایی S برگردانده می‌شود (شکل ۴). حال خروجی الگوریتم انتخاب نمونه که مجموعه داده‌ای نسبتاً متوازن است به طبقه‌بند داده می‌شود و معیارهای مد نظر بر روی مجموعه داده تست محاسبه می‌گردند.

1. Accuracy

۴- نتایج

در این بخش به ارزیابی کارایی روش پیشنهادی برای متعادل‌سازی مجموعه داده‌های نامتوازن پرداخته می‌شود. در ابتدا ابزار، مجموعه داده‌های نامتوازن مورد استفاده و تنظیمات پیاده‌سازی معرفی می‌گردد و سپس نتایج آزمایش‌های انجام‌گرفته ارائه می‌شود.

۴-۱ مجموعه داده‌های مورد استفاده، ابزار و تنظیمات پیاده‌سازی

جهت پیاده‌سازی روش پیشنهادی برای متعادل‌سازی مجموعه داده‌های نامتوازن از ابزار Jupyter notebook [۲۰] استفاده شده است. نرم‌افزار Jupyter notebook تحت زبان Python ۳.۶ اجرا می‌شود.

مجموعه داده‌های نامتوازن، یک مورد خاص برای مسئله طبقه‌بندی هستند که توزیع نمونه‌ها در آنها در بین کلاس‌ها یکنواخت نیست. این مجموعه‌ها معمولاً دو کلاس اکثریت (negative) و اقلیت (positive) دارند. روش پیشنهادی بر روی ۸ مجموعه داده نامتوازن انتخاب‌شده از مخزن داده KEEL[2] ارزیابی شده است. در این مقاله از مجموعه داده‌های دودویی Glass۶، Ecoli۳، Ecoli۰-۱-۳-۷-vs-۲-۶، Yeast۴، Yeast۳، Yeast۶، Abalone۱۹ و Page-blocks۰ استفاده گردیده که در دسته‌های متفاوتی از نسبت عدم توازن (IR)[3] قرار دارند؛ از مجموعه داده‌های نامتوازن با IR کم تا IR بسیار بالا. نسبت عدم توازن IR طبق (۸) محاسبه می‌گردد

$$IR = \frac{N_{maj}}{N_{\min}} \tag{۸}$$

که $N_{\min}$ تعداد نمونه‌های کلاس اقلیت و N_{maj} تعداد نمونه‌های کلاس اکثریت را نشان می‌دهد. هرچه نسبت عدم توازن (IR) به ۱ نزدیک‌تر باشد، آن مجموعه داده‌ای متعادل‌تر است. در جدول ۱ خلاصه‌ای از مشخصات مجموعه داده‌های مذکور آورده شده است.

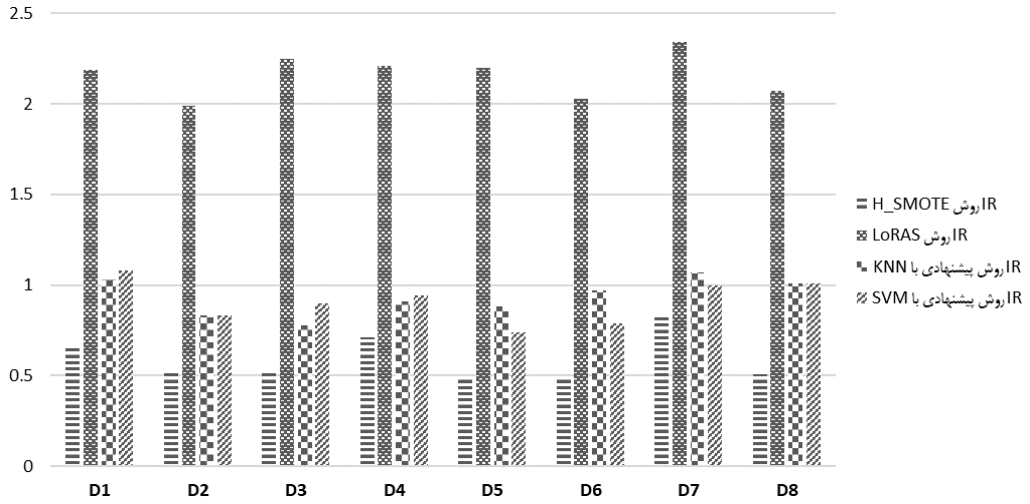
طبق جدول ۲ در روش LoRAS به k در صورتی که $|C_{\min}| \geq ۱۰۰$ باشد، به‌صورت پیش‌فرض مقدار ۳۰ داده می‌شود و در غیر این صورت مقدار ۵ می‌گیرد. اندازه N_{aff} برابر تعداد ویژگی‌هاست. مقادیر پارامترهای روش LoRAS بر اساس [۱۳] انتخاب شده است.

۴-۲ ارزیابی روش پیشنهادی برای متعادل‌سازی مجموعه داده‌های نامتوازن

روش پیشنهادی با الگوریتم‌های H-SMOTE [۱۲] و LoRAS [۱۳] مقایسه شده است. در دو الگوریتم مذکور تنها به افزایش تعداد نمونه‌های کلاس اقلیت پرداخته شده و همچنان عدم توازن زیادی در مجموعه داده‌های حاصل مشهود است. در روش پیشنهادی با انجام انتخاب نمونه بر روی ترکیب مجموعه داده‌های حاصل از دو الگوریتم، نسبت عدم توازن به‌طور قابل توجهی کاهش می‌یابد و توازن نسبی حاصل می‌گردد.

در شکل ۵، نسبت عدم توازن دو کلاس اقلیت و اکثریت برای روش پیشنهادی (با به‌کارگیری طبقه‌بندهای KNN و SVM) و دو الگوریتم داده‌افزایی مذکور آمده است. همان طور که مشاهده می‌شود، نسبت عدم توازن در روش پیشنهادی با هر دو طبقه‌بند نسبت به دو روش دیگر به ۱ نزدیک‌تر است. به‌عنوان مثال نسبت عدم توازن در مجموعه داده حاصل

2. https://sci2s.ugr.es/keel/datasets.php
3. Imbalance Rate



شکل ۵: نسبت عدم توازن (*IR*) مجموعه داده‌ها در روش پیشنهادی، H-SMOTE و LoRAS.

جدول ۱: مشخصات مجموعه داده‌های نامتوازن.

شناسه	مجموعه داده	تعداد نمونه‌ها	تعداد ویژگی‌ها	*IR*
مجموعه‌های داده‌ای با *IR* پایین (۱٫۵ تا ۹)				
D۱	Glass۶	۲۱۴	۹	۶٫۳۸
D۲	Yeast۳	۱۴۸۴	۸	۸٫۱
D۳	Page-blocks۰	۵۴۷۲	۱۰	۸٫۷۹
D۴	Ecoli۳	۳۳۶	۷	۸٫۶
مجموعه‌های داده‌ای با *IR* بالا (بیشتر از ۹)				
D۵	Yeast۴	۱۴۸۴	۸	۲۸٫۴۱
D۶	Yeast۶	۱۴۸۴	۸	۴۱٫۴
D۷	Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۲۸۱	۷	۳۹٫۱۴
D۸	Abalone۱۹	۴۱۷۴	۸	۱۲۹٫۴۴

جدول ۲: تنظیمات پارامترهای روش LoRAS.

مجموعه داده	تعداد نمونه‌های اقلیت	k	N_{aff}
Glass۶	۲۹	۵	۹
Yeast۳	۱۶۳	۳۰	۸
Page-blocks۰	۵۵۹	۳۰	۱۰
Ecoli۳	۳۵	۵	۷
Yeast۴	۵۱	۵	۸
Yeast۶	۳۵	۵	۸
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۷	۵	۷
Abalone۱۹	۳۲	۵	۸

از روش پیشنهادی با طبقه‌بند KNN برای مجموعه داده نامتوازن Yeast۴ به شناسه D۵ به میزان ۰٫۱۲ با مقدار استاندارد توازن فاصله دارد؛ در حالی که در روش H-SMOTE حدود ۰٫۵۱ با میزان توازن استاندارد فاصله دارد.

برای ارزیابی کارایی روش پیشنهادی برای مجموعه داده‌های نامتوازن، معیار صحت متعادل[1] طبقه‌بندهای KNN و SVM برای ۸ مجموعه داده‌ای نامتوازن ارزیابی گردیده است. صحت متعادل یکی از معیارهای شناخته‌شده در طبقه‌بندی مجموعه داده‌های نامتوازن است که بر اساس ماتریس درهم‌ریختگی[2] (شکل ۶) و مطابق (۹) محاسبه می‌شود

$$Balanced\ Accuracy = \frac{\frac{TP}{TP+FN} + \frac{TN}{TN+FP}}{۲} \qquad (۹)$$

اگر مجموعه داده کاملاً متعادل باشد– یعنی کلاس‌ها تقریباً یک اندازه باشند– صحت و صحت متعادل تمایل دارند به یک مقدار همگرا شوند. در واقع، تفاوت اصلی بین صحت متعادل و صحت، زمانی ظاهر می‌شود که مجموعه اولیه داده‌ها توزیع نامتعادلی را برای کلاس‌ها نشان دهد [۲۱].

مطابق با جدول ۳، صحت متعادل به‌دست‌آمده از دو طبقه‌بند KNN و SVM در روش پیشنهادی نسبت به سایر روش‌ها بهبود قابل ملاحظه‌ای داشته است. در این میان صحت متعادل به‌دست‌آمده از روش H-SMOTE برای مجموعه داده Page-blocks۰ با اختلاف اندکی (حدود ۰٫۰۰۴۸ برای KNN و ۰٫۰۱۱۵ برای SVM) نسبت به روش پیشنهادی بیشتر است؛ اما همچنان صحت متعادل روش پیشنهادی نسبت به روش LoRAS برای مجموعه داده Page-blocks۰ در هر دو طبقه‌بند بیشتر است.

یکی دیگر از معیارهای سنجش کارایی طبقه‌بندی مجموعه داده‌های نامتوازن، ویژگی است. معیار ویژگی به میزان توانایی یک طبقه‌بند برای یافتن نمونه‌های مثبت اشاره دارد. طبق (۱۰) برای به‌دست‌آوردن معیار ویژگی باید نسبت موارد منفی حقیقی را به مجموع موارد منفی حقیقی و مثبت کاذب حساب کرد

$$Specificity = \frac{TN}{TN+FP} \qquad (۱۰)$$

1. Balanced Accuracy
2. Confusion Matrix

		برچسب پیش‌بینی‌شده	
	کلاس‌ها	Positive (۱)	Negative (۰)
برچسب واقعی	Positive (۱)	*TP*	*FN*
	Negative (۰)	*FP*	*TN*

شکل ۶ ماتریس درهم‌ریختگی.

در جدول ۴، نتایج ارزیابی طبقه‌بندهای KNN و SVM برای مجموعه داده‌های نامتوازن بر اساس معیار ویژگی آمده است. بر اساس معیار ویژگی در اکثر مجموعه داده‌ها روش پیشنهادی به نتیجه بهتری نسبت به سایر روش‌ها دست یافته است. در مجموعه داده‌ای Page-blocks۰، معیار ویژگی روش H-SMOTE با اختلافی حدود ۰/۰۰۹۸ برای KNN و ۰/۰۱۲۴ برای SVM نتیجه بهتری را داشته است؛ اما همان طور که ملاحظه می‌شود معیار ویژگی روش پیشنهادی برای این مجموعه نسبت به روش LoRAS بیشتر است. در مجموعه داده‌ای Yeast۴ معیار ویژگی در روش LoRAS با طبقه‌بند SVM با اختلاف ۰/۰۳۳۵ نسبت به روش پیشنهادی بیشتر می‌باشد و در مجموعه داده‌ای Abalone۱۹ با طبقه‌بند KNN روش H-SMOTE و با طبقه‌بند SVM روش LoRAS عملکرد بهتری داشته است.

برای ارزیابی عملکرد طبقه‌بندهای در نظر گرفته شده، سایر معیارهای ارزیابی استاندارد مانند صحت، معیار F و AUC نیز اندازه‌گیری شدند. این معیارهای ارزیابی بر اساس ماتریس درهم‌ریختگی که در شکل ۶ نشان داده شده است، محاسبه می‌گردند. صحت یک طبقه‌بند نشان می‌دهد که چه میزان از نمونه‌های پیش‌بینی‌شده با برچسب واقعی کلاس مطابقت دارند. برای محاسبه صحت از (۱۱) استفاده می‌شود

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \qquad (۱۱)$$

معیار F بر اساس معیارهای دقت و فراخوانی محاسبه می‌شود؛ بنابراین به‌طور خلاصه به مفهوم معیارهای دقت و فراخوانی می‌پردازیم. معیار دقت برای یک طبقه‌بند با توجه به یک کلاس خاص، نسبت تعداد نمونه‌های به‌درستی پیش‌بینی‌شده به تعداد کل نمونه‌های آن کلاس است. فراخوانی برای یک کلاس خاص، نسبت نمونه‌هایی است که به‌درستی پیش‌بینی شده‌اند به تعداد کل نمونه‌هایی که واقعاً متعلق به کلاس هستند. معیارهای دقت و فراخوانی به‌ترتیب بر اساس (۱۲) و (۱۳) تعریف می‌شوند

$$precision = \frac{TP}{TP + FP} \qquad (۱۲)$$

$$recall = \frac{TP}{TP + FN} \qquad (۱۳)$$

معیار F نیز میانگین هارمونیک دقت و فراخوانی است که مطابق (۱۴) اندازه‌گیری می‌شود

$$F = \frac{۲ \times precision \times recall}{precision + recall} \qquad (۱۴)$$

منحنی (ROC) معیاری برای ارزیابی مسائل طبقه‌بندی دودویی است. ROC یک منحنی احتمال است که TPR را در برابر FPR[1] در آستانه‌های متفاوت ترسیم می‌کند. منحنی ROC را می‌توان با استفاده از مساحت زیر منحنی (AUC) برای تشخیص میزان توانایی طبقه‌بند برای تمایز بین کلاس‌ها خلاصه کرد. هرچه میزان AUC بالاتر باشد، عملکرد طبقه‌بند در تشخیص کلاس‌های اقلیت و اکثریت بهتر است. برای هر مجموعه داده، بالاترین مقدار صحت، F و AUC به‌دست‌آمده از هر طبقه‌بند در جداول ۵ و ۶ مشخص شده است.

همان طور که ذکر گردید جدول ۵ نتایج ارزیابی معیارهای صحت، F و AUC از طبقه‌بند KNN را بر روی ۸ مجموعه داده‌ای نشان می‌دهد. با توجه به نتایج ارائه‌شده از معیار صحت، روش پیشنهادی بر روی ۵ مجموعه داده‌ای از ۸ مجموعه داده عملکرد بهتری نسبت به سایر روش‌ها داشته و در سه مجموعه داده Page-blocks۰، Yeast۴ و Abalone۱۹ به‌ترتیب با اختلاف اندک ۰/۰۰۵۲، ۰/۰۰۲۵ و ۰/۰۰۰۳، H-SMOTE معیار صحت بهتری داشته است.

مطابق با نتایج به‌دست‌آمده از معیار F، روش پیشنهادی در اکثر مجموعه داده‌ها بهتر عمل کرده است و فقط در مجموعه داده‌ای Page-blocks۰ روش H-SMOTE با اختلاف ۰/۰۰۰۵ کارایی بیشتری داشته است. به همین ترتیب نتایج روش پیشنهادی بر اساس معیار AUC در اکثر مجموعه داده‌ها حکایت از عملکرد بهتر دارد؛ به غیر از مجموعه داده Page-blocks۰ که با اختلاف ۰/۰۰۴۸ روش H-SMOTE از نتیجه بهتری برخوردار بوده است.

مطابق با جدول ۶ با توجه به نتایج ارائه‌شده از طبقه‌بند SVM، روش پیشنهادی بر اساس معیار صحت در ۵ مجموعه داده از ۸ مجموعه نتیجه بهتری را نشان می‌دهد؛ به‌جز سه مجموعه داده Page-blocks۰، Yeast۴ و Abalone۱۹ که با روش H-SMOTE صحت بالاتری داشتند. معیار F روش پیشنهادی در اکثر مجموعه داده‌ها بالاتر است؛ به غیر از Page-blocks۰ و Abalone۱۹ که به ترتیب روش H-SMOTE و روش LoRAS نتیجه بهتری داشته است. بر اساس معیار AUC نیز روش پیشنهادی در ۷ مجموعه داده نتایج بهتری کسب کرده است؛ به‌جز در مجموعه داده Page-blocks۰ که روش H-SMOTE کارایی بیشتری داشته است.

در آزمایش دیگر برای ارزیابی اثربخشی اتوماتای یادگیر در انتخاب نمونه، روش مبتنی بر اتوماتای یادگیر با روش حذف تصادفی مقایسه شده است. در روش حذف تصادفی، پس از مرحله داده‌افزایی با H-SMOTE و LoRAS برای ایجاد تعادل در اندازه کلاس اقلیت و اکثریت، نمونه‌های کلاس C'_{min} به‌صورت کاملاً تصادفی حذف می‌شوند. در جداول ۷ و ۸ کارایی اتوماتای یادگیر و روش حذف تصادفی برای دو طبقه‌بند KNN و SVM مقایسه شده است. همان طور که مشاهده می‌گردد روش مبتنی بر اتوماتای یادگیر در تمامی مجموعه داده‌ها و برای هر دو طبقه‌بند KNN و SVM از صحت، F و AUC بالاتری برخوردار است. دلیل این امر، به‌کارگیری اتوماتای یادگیر برای شناسایی و حذف نمونه‌هایی است که تأثیر کمتری در طبقه‌بندی دارند. به بیانی دیگر، حذف تصادفی نمونه‌ها ممکن است منجر به حذف نمونه‌های اصلی و مؤثر در طبقه‌بندی و به تبع آن، کاهش کارایی طبقه‌بندی شود.

به‌طور خلاصه نتایج آزمایش‌های انجام‌گرفته نشان می‌دهند که روش پیشنهادی برای متعادل‌سازی مجموعه داده‌های نامتوازن در هر دو طبقه‌بند KNN و SVM در اکثر مجموعه داده‌های نامتوازن مورد مطالعه، عملکرد بهتری را در مقایسه با روش‌های H-SMOTE و LoRAS دارد. بر اساس نتایج به‌دست‌آمده برای روش پیشنهادی می‌توان اظهار کرد که برای طبقه‌بندی نمونه‌های مجموعه داده‌های نامتوازن به هر دو کلاس اقلیت و اکثریت، اهمیت یکسان داده می‌شود.

1. False Positive Rate

جدول ۳: نتایج ارزیابی طبقه‌بندهای KNN و SVM بر اساس معیار صحت متعادل.

مجموعه داده	H-SMOTE		LoRAS		روش پیشنهادی	
	SVM	KNN	SVM	KNN	SVM	KNN
Glass۶	۰/۹۷۷۵	۰/۹۶۰۰	۰/۹۶۲۷	۰/۹۴۴۲	۰/۹۸۹۷	۰/۹۸۰۷
Yeast۳	۰/۹۴۸۱	۰/۹۷۸۶	۰/۹۵۸۵	۰/۹۷۶۷	۰/۹۶۸۹	۰/۹۸۴۱
Page-blocks۰	۰/۹۷۸۰	۰/۹۷۹۹	۰/۹۴۵۲	۰/۹۵۲۱	۰/۹۶۶۵	۰/۹۷۵۱
Ecoli۳	۰/۹۱۵۶	۰/۹۴۸۲	۰/۹۱۶۸	۰/۹۶۲۴	۰/۹۵۵۱	۰/۹۷۷۸
Yeast۴	۰/۹۱۹۹	۰/۹۵۶۲	۰/۸۹۳۷	۰/۹۴۴۹	۰/۹۳۷۳	۰/۹۶۲۵
Yeast۶	۰/۹۵۱۴	۰/۹۷۸۵	۰/۹۴۹۰	۰/۹۷۳۹	۰/۹۷۲۶	۰/۹۸۷۱
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۷۱۲	۰/۹۸۸۵	۰/۹۱۰۶	۰/۹۶۳۴	۰/۹۸۸۵	۰/۹۹۳۵
Abalone۱۹	۰/۹۰۲۰	۰/۹۹۲۲	۰/۹۰۹۰	۰/۹۷۷۵	۰/۹۱۷۰	۰/۹۹۲۵

جدول ۴: نتایج ارزیابی طبقه‌بندهای KNN و SVM بر اساس معیار ویژگی.

مجموعه داده	H-SMOTE		LoRAS		روش پیشنهادی	
	SVM	KNN	SVM	KNN	SVM	KNN
Glass۶	۰/۹۶۷۷	۰/۹۳۲۲	۰/۹۸۴۳	۰/۹۲۸۵	۱	۱
Yeast۳	۰/۹۰۱۸	۰/۹۶۶۸	۰/۹۳۷۱	۰/۹۶۹۳	۰/۹۴۰۱	۰/۹۷۶۹
Page-blocks۰	۰/۹۷۰۱	۰/۹۷۴۸	۰/۹۳۹۵	۰/۹۴۸۶	۰/۹۵۷۷	۰/۹۶۵۰
Ecoli۳	۰/۸۳۹۰	۰/۸۹۶۵	۰/۸۵۸۶	۰/۹۲۴۷	۰/۹۲۰۴	۰/۹۶۵۵
Yeast۴	۰/۸۴۵۷	۰/۹۱۹۵	۰/۹۲۸۲	۰/۹۲۲۸	۰/۸۹۴۷	۰/۹۳۱۲
Yeast۶	۰/۹۰۹۷	۰/۹۵۹۴	۰/۹۴۶۱	۰/۹۵۷۱	۰/۹۴۹۳	۰/۹۷۶۶
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۶۲۵	۰/۹۷۷۰	۰/۸۸۳۷	۰/۹۲۶۸	۰/۹۷۷۰	۱
Abalone۱۹	۰/۸۰۵۲	۰/۹۹۱۲	۰/۸۶۹۸	۰/۹۷۱۸	۰/۸۴۱۲	۰/۹۸۸۴

جدول ۵: نتایج ارزیابی طبقه‌بند KNN بر اساس معیارهای صحت، F و AUC.

مجموعه داده	H-SMOTE			LoRAS			روش پیشنهادی		
	صحت	F	AUC	صحت	F	AUC	صحت	F	AUC
Glass۶	۰/۹۶۴۵	۰/۹۵۶۵	۰/۹۶۰۰	۰/۹۳۸۲	۰/۹۵۴۱	۰/۹۴۴۲	۰/۹۸۱۶	۰/۹۸۲۷	۰/۹۸۰۷
Yeast۳	۰/۹۸۲۱	۰/۹۷۴۲	۰/۹۷۸۶	۰/۹۷۴۲	۰/۹۸۰۶	۰/۹۷۶۷	۰/۹۸۴۶	۰/۹۸۳۲	۰/۹۸۴۱
Page-blocks۰	۰/۹۸۱۵	۰/۹۷۳۲	۰/۹۷۹۹	۰/۹۵۰۸	۰/۹۶۶۳	۰/۹۵۲۱	۰/۹۷۶۳	۰/۹۷۲۷	۰/۹۷۵۱
Ecoli۳	۰/۹۵۸۳	۰/۹۴۵۴	۰/۹۴۸۲	۰/۹۴۶۹	۰/۹۶۰۸	۰/۹۶۲۴	۰/۹۷۸۸	۰/۹۷۶۷	۰/۹۷۷۸
Yeast۴	۰/۹۶۷۹	۰/۹۵۱۲	۰/۹۵۶۲	۰/۹۳۶۰	۰/۹۵۲۹	۰/۹۴۴۹	۰/۹۶۵۴	۰/۹۶۰۷	۰/۹۶۲۵
Yeast۶	۰/۹۸۵۳	۰/۹۷۶۳	۰/۹۷۸۵	۰/۹۶۸۵	۰/۹۷۵۷	۰/۹۷۳۹	۰/۹۸۷۲	۰/۹۸۷۰	۰/۹۸۷۱
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۸۸۸	۰/۹۸۸۰	۰/۹۸۸۵	۰/۹۴۹۱	۰/۹۶۲۹	۰/۹۶۳۴	۰/۹۹۳۷	۰/۹۹۳۴	۰/۹۹۳۵
Abalone۱۹	۰/۹۹۲۶	۰/۹۹۰۵	۰/۹۹۲۲	۰/۹۷۵۵	۰/۹۸۲۲	۰/۹۷۷۵	۰/۹۹۲۳	۰/۹۹۲۶	۰/۹۹۲۵

جدول ۶: نتایج ارزیابی طبقه‌بند SVM بر اساس معیارهای صحت، F و AUC.

مجموعه داده	H-SMOTE			LoRAS			روش پیشنهادی		
	صحت	F	AUC	صحت	F	AUC	صحت	F	AUC
Glass۶	۰/۹۷۸۷	۰/۹۷۵۶	۰/۹۷۷۵	۰/۹۷۵۳	۰/۹۸۴۳	۰/۹۶۲۷	۰/۹۹۰۵	۰/۹۹۱۳	۰/۹۸۹۷
Yeast۳	۰/۹۶۲۵	۰/۹۴۳۱	۰/۹۴۸۱	۰/۹۵۱۸	۰/۹۶۲۳	۰/۹۵۸۵	۰/۹۷۲۹	۰/۹۶۷۷	۰/۹۶۸۹
Page-blocks۰	۰/۹۸۰۵	۰/۹۷۱۸	۰/۹۷۸۰	۰/۹۳۹۵	۰/۹۵۹۲	۰/۹۴۵۲	۰/۹۶۶۹	۰/۹۶۴۹	۰/۹۶۶۵
Ecoli۳	۰/۹۳۰۵	۰/۹۰۶۸	۰/۹۱۵۶	۰/۸۹۳۹	۰/۹۱۸۶	۰/۹۱۶۸	۰/۹۵۶۹	۰/۹۵۲۹	۰/۹۵۵۱
Yeast۴	۰/۹۴۴۵	۰/۹۱۰۶	۰/۹۱۹۹	۰/۹۰۴۹	۰/۹۲۶۵	۰/۸۹۳۷	۰/۹۴۲۳	۰/۹۳۲۰	۰/۹۳۷۳
Yeast۶	۰/۹۶۵۹	۰/۹۴۵۶	۰/۹۵۱۴	۰/۹۴۸۰	۰/۹۶۰۷	۰/۹۴۹۰	۰/۹۷۳۹	۰/۹۷۱۸	۰/۹۷۲۶
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۷۲۲	۰/۹۶۶۴	۰/۹۷۱۲	۰/۸۹۸۳	۰/۹۲۶۸	۰/۹۱۰۶	۰/۹۸۷۸	۰/۹۸۶۳	۰/۹۸۸۵
Abalone۱۹	۰/۹۳۲۹	۰/۸۹۰۹	۰/۹۰۲۰	۰/۸۹۵۲	۰/۹۱۸۱	۰/۹۰۹۰	۰/۹۱۸۴	۰/۹۰۹۴	۰/۹۱۷۰

۵- نتیجه‌گیری

عدم توازن کلاس‌ها در مجموعه داده‌ها باعث نادیده‌گرفته‌شدن کلاسی می‌شود که تعداد نمونه‌های کمتری دارد (کلاس اقلیت) و این در حالی است که در اغلب مسائل طبقه‌بندی این کلاس از اهمیت بیشتری برخوردار است. جهت رفع این مشکل بایست به متعادل‌نمودن این گونه

جدول ۷: مقایسه کارایی روش مبتنی بر اتوماتای یادگیر با روش حذف تصادفی برای طبقه‌بند KNN.

مجموعه داده	حذف تصادفی			اتوماتای یادگیر		
	صحت	*F*	*AUC*	صحت	*F*	*AUC*
Glass۶	۰/۹۴۸۷	۰/۹۴۳۳	۰/۹۴۸۲	۰/۹۸۱۶	۰/۹۸۲۷	۰/۹۸۰۷
Yeast۳	۰/۹۷۳۳	۰/۹۷۳۵	۰/۹۷۳۵	۰/۹۸۴۶	۰/۹۸۳۲	۰/۹۸۴۱
Page-blocks۰	۰/۹۶۹۲	۰/۹۶۸۹	۰/۹۶۹۲	۰/۹۷۶۳	۰/۹۷۲۷	۰/۹۷۵۱
Ecoli۳	۰/۹۲۶۷	۰/۹۱۵۶	۰/۹۲۲۰	۰/۹۷۸۸	۰/۹۷۶۷	۰/۹۷۷۸
Yeast۴	۰/۹۵۲۸	۰/۹۴۳۰	۰/۹۴۷۲	۰/۹۶۵۴	۰/۹۶۰۷	۰/۹۶۲۵
Yeast۶	۰/۹۷۳۷	۰/۹۷۲۳	۰/۹۷۳۱	۰/۹۸۷۲	۰/۹۸۷۰	۰/۹۸۷۱
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۶۹۶	۰/۹۷۱۴	۰/۹۷۲۲	۰/۹۹۳۷	۰/۹۹۳۴	۰/۹۹۳۵
Abalone۱۹	۰/۹۹۰۲	۰/۸۹۹۹	۰/۹۹۰۱	۰/۹۹۲۳	۰/۹۹۲۶	۰/۹۹۲۵

جدول ۸: مقایسه کارایی روش مبتنی بر اتوماتای یادگیر با روش حذف تصادفی برای طبقه‌بند SVM.

مجموعه داده	حذف تصادفی			اتوماتای یادگیر		
	صحت	*F*	*AUC*	صحت	*F*	*AUC*
Glass۶	۰/۹۵۷۲	۰/۹۵۲۳	۰/۹۵۶۰	۰/۹۹۰۵	۰/۹۹۱۳	۰/۹۸۹۷
Yeast۳	۰/۹۵۴۳	۰/۹۵۳۸	۰/۹۵۴۹	۰/۹۷۲۹	۰/۹۶۷۷	۰/۹۶۸۹
Page-blocks۰	۰/۹۶۲۴	۰/۹۶۲۷	۰/۹۶۲۳	۰/۹۶۶۹	۰/۹۶۴۹	۰/۹۶۶۵
Ecoli۳	۰/۹۱۰۹	۰/۸۹۴۴	۰/۹۰۴۴	۰/۹۵۶۹	۰/۹۵۲۹	۰/۹۵۵۱
Yeast۴	۰/۹۲۹۳	۰/۹۱۴۰	۰/۹۲۲۵	۰/۹۴۲۳	۰/۹۳۲۰	۰/۹۳۷۳
Yeast۶	۰/۹۵۴۳	۰/۹۵۰۹	۰/۹۵۳۱	۰/۹۷۳۹	۰/۹۷۱۸	۰/۹۷۲۶
Ecoli۰ – ۱ – ۳ – ۷ – vs – ۲ – ۶	۰/۹۶۳۶	۰/۹۶۵۵	۰/۹۶۶۶	۰/۹۸۷۸	۰/۹۸۶۳	۰/۹۸۸۵
Abalone۱۹	۰/۹۱۰۱	۰/۸۹۹۰	۰/۹۰۷۹	۰/۹۱۸۴	۰/۹۰۹۴	۰/۹۱۷۰

مجموعه داده‌ها پرداخت.

در این مقاله مشکل عدم توازن این گونه مجموعه داده‌ها، بررسی و در این راستا، روشی جدید برای متعادل‌نمودن مجموعه داده‌های نامتوازن پیشنهاد گردید. در روش پیشنهادی، ابتدا نمونه‌های کلاس اقلیت در هر مجموعه داده‌ای نامتوازن با دو روش H-SMOTE و LoRAS داده‌افزایی شدند و با توجه به اینکه نسبت عدم توازن مجموعه داده‌ها همچنان بالا بود، روشی ابتکاری برای رسیدن به تعادل نسبی ارائه گردید. در این روش پس از داده‌افزایی، عمل انتخاب نمونه بر روی نمونه‌های کلاس اقلیت پیاده می‌شود. در عمل انتخاب نمونه به هر نمونه از کلاس اقلیت، یک اتوماتای یادگیر جهت انتخاب‌شدن یا نشدن در مجموعه کاهش‌یافته نهایی نسبت داده می‌شود و طبق نتیجه صحت طبقه‌بندهای SVM و KNN به اتوماتای یادگیر پاداش یا جریمه داده می‌شود.

بر اساس نتایج آزمایش‌های انجام‌گرفته در روش پیشنهادی نسبت به روش‌های H-SMOTE و LoRAS، مجموعه داده‌ها از نسبت عدم توازن مناسب‌تری برخوردار هستند. همچنین در مجموع، روش پیشنهادی بر اساس معیارهای رایج طبقه‌بندی مجموعه داده‌های نامتوازن یعنی صحت متعادل و ویژگی در مقایسه با دو روش مذکور عملکرد بهتری داشته است.

مراجع

[1] H. Kim, H. Cho, and D. Ryu, "Corporate bankruptcy prediction using machine learning methodologies with a focus on sequential data," *Computational Economics*, vol. 59, pp. 1231-1249, 2022.

[2] D. Yousif Mikhail, F. Al-Mukhtar, and S. Wahab Kareem, "A comparative evaluation of cancer classification via TP53 gene mutations using machine learning," *Asian Pacific J. of Cancer Prevention*, vol. 23, no. 7, pp. 2459-2467, Jul. 2022.

[3] L. Yang and Y. Jiachen, "Few-shot cotton pest recognition and terminal," *Computers and Electronics in Agriculture*, vol. 169, Article ID: 105240, 2020.

[4] P. Kumar, R. Bhatnagar, K. Gaur, and A. Bhatnagar, "Classification of imbalanced data: review of methods and applications," *IOP Conf. Series: Materials Science and Engineering*, vol. 1099, no 1, Article ID: 012077, 2021.

[5] C. F. Tsai, W. C. Lin, Y. H. Hu, and G. T. Yao, "Under-sampling class imbalanced datasets by combining clustering analysis and instance selection," *Information Sciences*, vol. 477, pp. 47-54, Mar. 2019.

[6] I. Czarnowski and P. Jedrzejowicz, "An approach to imbalanced data classification based on instance selection and over-sampling," in *Proc. 11th Int. Conf.on Computational Collective Intelligence,* pp. 601-610, Hendaye, France, 4-6 Sept. 2019.

[7] D. Gan, J. Shen, B. An, M. Xu, and N. Liu, "Integrating TANBN with cost sensitive classification algorithm for imbalanced data in medical diagnosis," *Computers & Industrial Engineering*, vol. 140, Article ID: 106266, Feb. 2020.

[8] L. Yang and Y. Jiachen, "Meta-learning baselines and database for few-shot classification in agriculture," *Computers and Electronics in Agriculture*, vol. 182, Article ID: 106055, Mar. 2021.

[9] Z. Peng, Z. Li, J. Zhang, Y. Li, G. J. Qi, and J. Tang, "Few-shot image recognition with knowledge transfer," in *Proc. of the IEEE/CVF Int. Conf. on Computer Vision*, pp. 441-449, Seoul, South Korea, 27 Oct.-2 Nov. 2019.

[10] F. Jimenez, G. Sanchez, J. Palma, and G. Sciavicco, "Three-objective constrained evolutionary instance selection for classification: wrapper and filter approaches," *Engineering Applications of Artificial Intelligence*, vol. 107, Article ID: 104531, Jan. 2022.

[11] G. E. Melo-Acosta, F. Duitama-Muñoz, and J. D. Arias-Londoño, *An Instance Selection Algorithm for Big Data in High Imbalanced Datasets Based on LSH*, arXiv: 2210.04310, Oct. 2022.

[12] X. Chao and L. Zhang, "Few-shot imbalanced classification based on data augmentation," *Multimedia Systems*, vol. 29, no. 5, pp. 2843-2851, 2023.

[13] S. Bej, N. Davtyan, M. Wolfien, M. Nassar, and O. Wolkenhauer, "LoRas: an oversampling approach for imbalanced datasets," *Machine Learning*, vol. 110, pp. 279-301, 2021.

[14] J. C. Requelme, J. S. Aguilar-Ruiz, and M. Toro, "Finding representative patterns with ordered projections," *Pattern Recognition*, vol. 36, no. 4, pp. 1009-1018, Apr. 2003.

[15] D. R. Wilson and T. R. Martinez, "Instance pruning techniques," in *Proc. of the 14th Int. Conf. on Machine Learning*, pp. 400-411, 8-12 Jul. 1997.

[16] M. Moran, T. Cohen, Y. Ben-Zion, and G. Gordon, "Curious instance selection," *Information Sciences*, vol. 608, pp. 794-808, Aug. 2022.

[17] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *J. of Artificial Intelligence Research*, vol. 16, pp. 321-357, Jan. 2002.

[۱۸] ش. سرگلزایی، ف. حسین‌زاده سلجوقی و ه. آقایاری، "ارائه روشی نوین برای رتبه‌بندی اعداد فازی با استفاده از مرکز محیطی دایره و کاربرد آن در ارزیابی عملکرد مدیریت زنجیره تأمین،" *نشریه تصمیم‌گیری و تحقیق در عملیات*، دوره ۳، شماره ۳، صص. ۲۳۶-۲۴۸، پاییز ۱۳۹۷.

[19] S. N. Kumpati and A. T. Mandayam, *Learning Automata: An Introduction*, Courier Corporation, 2012.

[20] J. C. Dominguz, *et al.*, "Teaching chemical engeering using Jupyter notebook: problem generators and lecturing tools," *Education for Chemical Engineers*, vol. 37, pp. 1-10, Oct. 2021.

[21] M. Grandini, E. Bagli, and G. Visani, *Multi-Class Classification: An Overview,* arXiv:2008.05756, Aug. 2020.

پرستو محقق مدرک کارشناسی خود را در رشته مهندسی کامپیوتر گرایش فناوری اطلاعات در سال ۱۳۹۸ از دانشگاه زابل دریافت کرد و در حال حاضر دانشجوی کارشناسی ارشد در رشته مهندسی فناوری اطلاعات گرایش مدیریت سیستم‌های اطلاعاتی در دانشگاه سیستان و بلوچستان است. زمینه‌های تحقیقاتی مورد علاقه ایشان عبارتند از مدیریت سیستم‌های اطلاعاتی، علم داده، یادگیری ماشین و طبقه‌بندی.

سمیرا نوفرستی تحصیلات خود را در مقاطع کارشناسی و کارشناسی ارشد مهندسی کامپیوتر به‌ترتیب در سال‌های ۱۳۸۲ و ۱۳۸۴ از دانشگاه صنعتی شریف و دانشگاه صنعتی امیرکبیر و در مقطع دکتری مهندسی کامپیوتر در سال ۱۳۹۴ از دانشگاه شهید بهشتی به پایان رساند و هم‌اکنون استادیار دانشکده مهندسی برق و کامپیوتر دانشگاه سیستان و بلوچستان است. زمینه‌های تحقیقاتی اصلی مورد علاقه ایشان عبارتند از هوش مصنوعی، پردازش زبان طبیعی، متن‌کاوی و تحلیل احساسات.

مهری رجائی در سال ۱۳۸۲ مدرک کارشناسی مهندسی کامپیوتر خود را از دانشگاه صنعتی شریف و در سال ۱۳۸۴ مدرک کارشناسی ارشد مهندسی کامپیوتر خود را از دانشگاه صنعتی امیرکبیر دریافت نمود. در سال ۱۳۹۴ موفق به اخذ درجه دکترا در رشته مهندسی کامپیوتر از دانشگاه علم و صنعت شد. وی از سال ۱۳۸۴ در دانشکده مهندسی برق و کامپیوتر دانشگاه سیستان و بلوچستان مشغول به فعالیت گردید و اینک نیز عضو هیأت علمی این دانشکده است. زمینه‌های علمی مورد علاقه نام‌برده شامل شبکه‌های اجتماعی، حفظ حریم خصوصی در انتشار شبکه‌های اجتماعی، پایگاه داده و محاسبات نرم است.