

Semi-Supervised Classification of Imbalanced Data Streams Under Concept Drift Using Chaotic Sequences

Faezeh Azadi¹, Javad Hamidzadeh^{2*}

1, 2- Faculty of Computer Engineering and Information Technology, Sadjad University, Mashhad, Iran.

¹ fae.azadi531@sadjad.ac.ir, ^{2*} j_hamidzadeh@sadjad.ac.ir

Corresponding author address: Javad Hamidzadeh, Faculty of Computer Engineering and Information Technology, Sadjad University, No. 64 Jalal Al Ahmad St, Mashhad, Iran, Post Code: 9188148848.

Abstract- Real-world data stream classification remains a fundamental challenge in machine learning due to the simultaneous occurrence of concept drift, class imbalance, and the limited availability of labeled data. Existing semi-supervised learning methods typically address only one of these challenges. In this study, a semi-supervised active learning framework based on a Radial Basis Function Neural Network (RBFNN) is proposed for data stream classification under simultaneous concept drift and class imbalance. In the proposed framework, a sliding window mechanism is employed to retain newly arrived samples while gradually discarding outdated data, enabling the model to remain consistent with the current data distribution. Subsequently, an active learning strategy based on the combination of uncertainty and density criteria selects the most informative samples for labeling, thereby reducing labeling costs while improving training quality. Furthermore, a smooth pseudo-labeling mechanism based on feature-space similarity is adopted to effectively exploit unlabeled data. In the modeling stage, the RBFNN architecture is enhanced with an adaptive center adjustment mechanism and dynamic updating to cope with concept drift. The initial RBF centers are initialized using an improved K-means++ algorithm and are subsequently updated online according to incoming data. In addition, to address class imbalance, a synthetic sample generation method based on weighted combinations of minority-class samples and the Logistic Map chaotic sequence is proposed, increasing the diversity of generated samples. Moreover, a weighted Focal Loss function is employed to improve the model's sensitivity to difficult samples while dynamically adjusting class weights. The network parameters are optimized using the Adam algorithm. Experimental results demonstrate that the integration of active learning, adaptive RBF center adjustment, semi-supervised pseudo-labeling, and mechanisms for handling concept drift and class imbalance significantly improves the performance of data stream classification. Compared with state-of-the-art methods, including SSALD, SSALE-EIS, AACE-DI, LEPID, and CBAL, the proposed method achieves superior performance, with average improvements of 4% in F1-score, 3% in G-Mean, and 2.5% in Specificity. These results confirm the effectiveness and robustness of the proposed framework for real-world data stream classification under concept drift and class imbalance.

Keywords- Semi-Supervised learning, Data Stream, Imbalanced Data, Concept Drift.

طبقه‌بندی نیمه‌نظارتی جریان داده‌های نامتوازن تحت رانش مفهوم با استفاده از دنباله آشوبی

فائزه آزادی^۱، جواد حمیدزاده^{۲*}

۱ و ۲- دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد، ایران

¹ fae.azadi531@sadjad.ac.ir, ² j_hamidzadeh@sadjad.ac.ir

نشانی نویسنده مسئول: جواد حمیدزاده، مشهد، بلوار جلال آل احمد، جلال آل احمد ۶۴، دانشگاه سجاد، دانشکده مهندسی کامپیوتر و فناوری اطلاعات، کد پستی: ۹۱۸۸۱۴۸۸۴۸

چکیده- طبقه‌بندی جریان‌های داده‌ای در دنیای واقعی، به دلیل بروز هم‌زمان رانش مفهوم، داده‌های نامتوازن و محدودیت در دسترسی به داده‌های برچسب‌خورده، همچنان یکی از چالش‌های اساسی یادگیری ماشین است. روش‌های موجود در یادگیری نیمه‌نظارتی معمولاً تنها یکی از این چالش‌ها را پوشش داده‌اند. در این پژوهش، یک چارچوب یادگیری فعال نیمه‌نظارتی مبتنی بر شبکه عصبی تابع پایه شعاعی (RBFNN) برای طبقه‌بندی جریان‌های داده‌ای در حضور هم‌زمان رانش مفهوم و داده‌های نامتوازن ارائه شده است. در چارچوب پیشنهادی، از مکانیزم پنجره لغزان برای نگهداری داده‌های جدید و حذف تدریجی داده‌های قدیمی استفاده می‌شود تا مدل با توزیع جاری داده‌ها سازگار بماند. سپس، راهبرد یادگیری فعال مبتنی بر ترکیب معیارهای عدم قطعیت و چگالی، نمونه‌های دارای بیشترین ارزش اطلاعاتی را برای برچسب‌گذاری انتخاب می‌کند و ضمن کاهش هزینه برچسب‌گذاری، کیفیت آموزش را افزایش می‌دهد. همچنین، برای بهره‌گیری از داده‌های بدون برچسب، از مکانیزم برچسب‌گذاری هموار مبتنی بر شباهت در فضای ویژگی استفاده شده است. در بخش مدل‌سازی، ساختار RBFNN با مکانیزم تنظیم تطبیقی مراکز و به‌روزرسانی پویا برای سازگاری با رانش مفهوم توسعه یافته است. مراکز اولیه نورون‌های RBF با نسخه بهبودیافته الگوریتم K-means++ مقداردهی اولیه شده و سپس متناسب با داده‌های جدید به صورت آنلاین بازتنظیم می‌شوند. علاوه بر این، برای مقابله با داده‌های نامتوازن، روشی برای تولید نمونه‌های مصنوعی مبتنی بر ترکیب وزن‌دار نمونه‌های کلاس اقلیت و دنباله آشوبی Logistic Map ارائه شده است که تنوع نمونه‌های تولیدشده را افزایش می‌دهد. همچنین، از تابع خطای Focal Loss وزن‌دار برای افزایش حساسیت مدل نسبت به نمونه‌های دشوار و تنظیم پویای وزن کلاس‌ها استفاده شده و پارامترهای شبکه با الگوریتم Adam بهینه‌سازی می‌شوند. نتایج ارزیابی نشان می‌دهد که ترکیب یادگیری فعال، تنظیم تطبیقی مراکز RBF، برچسب‌گذاری نیمه‌نظارتی و مکانیزم‌های مقابله با رانش مفهوم و عدم تعادل داده‌ها، عملکرد مدل را در طبقه‌بندی جریان‌های داده‌ای بهبود می‌دهد. همچنین، روش پیشنهادی در مقایسه با روش‌های SSALD، SSALE-EIS، AACE-LEPID، DI و CBAL عملکرد بهتری داشته و به‌طور میانگین ۴ درصد بهبود در معیار F1-score، ۳ درصد بهبود در معیار G-Mean و ۲/۵ درصد بهبود در معیار Specificity به دست آورده است که بیانگر کارایی و پایداری مناسب چارچوب پیشنهادی در شرایط واقعی است.

واژه‌های کلیدی: یادگیری نیمه نظارتی، جریان داده، داده‌های نامتعادل، رانش مفهوم.

۱- مقدمه

در سال‌های اخیر، یادگیری از جریان‌های داده به یکی از موضوعات مهم و پرچالش در حوزه‌ی هوش مصنوعی و داده‌کاوی تبدیل شده است. منظور از «جریان داده»، ورود پیوسته و ترتیبی نمونه‌های داده در طول زمان است؛ به گونه‌ای که داده‌ها به صورت مرحله‌ای و آنلاین در اختیار مدل قرار می‌گیرند و امکان دسترسی هم‌زمان به تمامی داده‌ها وجود ندارد. در چنین محیطی، مدل باید به صورت تدریجی و افزایشی، هم‌زمان با ورود نمونه‌های جدید، به روزرسانی شود. همچنین، در دیتاست‌های مورد استفاده، داده‌ها به صورت یک توالی زمانی یا ترتیبی پردازش شده‌اند؛ بدین معنا که نمونه‌ها به ترتیب ورود به سیستم به مدل ارائه می‌شوند و توزیع داده‌ها ممکن است در طول زمان تغییر کند. در بسیاری از کاربردهای دنیای واقعی، مانند شناسایی بدافزارها [۱] و تحلیل داده‌های شبکه‌های اجتماعی [۲]، حجم عظیمی از داده‌ها به صورت پیوسته و متوالی تولید می‌شود. به این نوع داده‌ها، جریان داده گفته می‌شود. در بهبود معیار دقت در یادگیری نیمه‌نظارتی جریان‌های داده با داده‌های نامتوازن و رانش مفهوم در طبقه‌بندی جریان‌های داده، چالش‌های متعددی وجود دارد که مانع از عملکرد مطلوب روش‌های سنتی می‌شود.

یکی از چالش‌های اصلی، رانش مفهوم در نظر گرفته می‌شود. رانش مفهوم به تغییر توزیع داده‌ها در طول زمان به دلیل تغییر در محیط یا شرایط کاری اطلاق می‌شود. بدون از دست دادن کلیت، طبقه‌بندی‌کننده‌های ایجادشده بر اساس داده‌های تاریخی ممکن است در تشخیص دقیق نمونه‌های تازه‌واردشده دچار شکست شوند و در نتیجه عملکرد ضعیفی داشته شود [۳].

چالش دیگر، داده‌های نامتوازن داده‌ها مطرح می‌شود. به کلاس‌هایی که تعداد داده‌های بسیار زیادی دارند، کلاس اکثریت گفته می‌شود و به کلاس‌هایی که داده‌های بسیار کمی را شامل می‌شوند، کلاس اقلیت اطلاق می‌شود. معمولاً کلاس اقلیت از اهمیت بالایی برخوردار دانسته می‌شود و در بیشتر مسائل، شناسایی داده‌های این کلاس‌ها به عنوان هدف اصلی در نظر گرفته می‌شود. مواجهه با مسائل کلاس‌های نامتوازن کاری بسیار دشوار تلقی می‌شود و از سوی دیگر، بیشتر مسائل دنیای واقعی دارای کلاس‌هایی با توزیع غیریکنواخت هستند. بنابراین، در ارائه یک روش کاهش نمونه، نحوه توزیع داده‌ها در بین کلاس‌ها به عنوان یک چالش مهم باید مدنظر قرار داده شود. موضوع مهم در کلاس‌های نامتوازن این است که در نظر گرفتن معیار صحت طبقه‌بندی به تنهایی کافی دانسته نمی‌شود، زیرا طبقه‌بندی

نادرست نمونه‌های کلاس اقلیت تأثیر کمی در نرخ صحت ایجاد می‌کند. به این معنا که نرخ صحت بالا ممکن است تنها از نمونه‌های کلاس اکثریت ناشی شود. از این رو، در مواجهه با مسائلی که دارای کلاس‌های نامتوازن هستند، اغلب معیارهای دیگری نیز در نظر گرفته می‌شود [۴]. این موضوع باعث می‌شود که تمایل مدل به پیش‌بینی کلاس‌های پرتکرار افزایش داده شود و تشخیص صحیح کلاس‌های کم‌تکرار به درستی انجام نشود. چنین وضعیتی منجر به ایجاد سوگیری در مدل و کاهش توانایی آن در شناسایی رخداد‌های نادر اما مهم می‌شود.

سومین چالش، کمبود داده‌های برچسب‌دار شناخته می‌شود. یک جریان داده ممکن است شامل نمونه‌های بالقوه بی‌نهایت باشد. برچسب‌گذاری دستی تمامی آن‌ها دشوار و پرهزینه محسوب می‌شود و در نتیجه، تعداد نمونه‌های برچسب‌گذاری شده محدود می‌شود. بنابراین، در بیشتر موارد، اطلاعات پایه یک نمونه ورودی ممکن است در دسترس نباشد [۵]. در مجموع، وجود رانش مفهوم، نامتوازن بودن کلاس‌ها و کمبود برچسب‌ها موجب می‌شود که روش‌های سنتی یادگیری که برای داده‌های ایستا و کاملاً برچسب‌دار طراحی شده‌اند، در محیط‌های پویا و جریانی کارایی خود را از دست بدهند. از این رو، توسعه روش‌های سازگار و هوشمند که بتوانند با این شرایط پویا و محدودیت‌های برچسب‌گذاری مقابله کنند، ضروری دانسته می‌شود.

در یک سناریوی واقعی از تشخیص عیب در اجزای صنعتی مانند یاتاقان‌ها، چرخ‌دنده‌ها، شفت‌ها و پمپ‌ها هر سه چالش رانش مفهوم، داده‌های نامتوازن داده‌ها و محدودیت برچسب‌گذاری مشاهده می‌شود [۶، ۷]. در شرایط کاری که به مرور زمان تغییر می‌کنند، ویژگی‌های این اجزا دچار نوسان می‌شوند و در نتیجه توزیع داده‌های مربوط به آن‌ها به صورت پویا تغییر داده می‌شود. این پدیده به عنوان رانش مفهوم شناخته می‌شود [۸، ۹]. از سوی دیگر، خرابی تجهیزات معمولاً کمتر از حالت عادی رخ می‌دهد؛ بنابراین داده‌های مربوط به خرابی در مقایسه با داده‌های حالت عادی بسیار کمتر بوده و داده‌های نامتوازن داده‌ها ایجاد می‌شود [۲، ۱۰]. همچنین، از آنجا که وضعیت اجزا به طور پیوسته پایش می‌شود، برچسب‌گذاری دستی تمام داده‌های نظارتی امری دشوار و زمان‌بر تلقی می‌شود.

نوآوری‌های اصلی این پژوهش در پنج محور کلیدی و در قالب یک چارچوب یکپارچه ارائه شده است: شبکه عصبی RBFNN تطبیقی با استفاده از K-means++ برای تنظیم پویای مراکز و ارزیابی نمایندگی داده‌ها، به صورت پویا با تغییرات توزیع داده و رانش

بررسی چالش رانش مفهوم: در این گروه پژوهشی صرفاً به بررسی چالش رانش مفهوم پرداخته شده است. عدم توجه به داده‌های نامتوازن و محدودیت برچسب‌گذاری به‌عنوان یک ضعف قابل توجه در این پژوهش‌ها تلقی می‌شود و نشان‌دهنده نیاز به رویکردهایی است که بتوانند تمامی این محدودیت‌ها و چالش‌ها را به‌صورت یکپارچه مدیریت کنند. در مطالعه [۱۱]، الگوریتم OEC^1 با استفاده از یک ساختار دو لایه به‌کار گرفته شده است. در لایه داخلی، یک نمونه ورودی توسط هر طبقه‌بند پایه پیش‌بینی شده و بر اساس برچسب واقعی، خطای مربوطه محاسبه و به‌روزرسانی انجام می‌شود. در لایه بیرونی، وزن هر طبقه‌بند به‌صورت پویا و با استفاده از الگوریتم NEG^2 تنظیم می‌شود تا مجموع ضرر مورد انتظار کاهش داده شود. خروجی نهایی بر اساس ترکیب وزن‌ها و پیش‌بینی‌های طبقه‌بندها تعیین می‌شود که این ساختار موجب افزایش دقت و مقاومت مدل در برابر رانش مفهوم و نویز می‌شود.

در مطالعه [۱۲]، رویکردی با ترکیب یادگیری گروهی، یادگیری افزایشی و پنجره لغزان برای مقابله با رانش مفهوم طراحی شده است. مدل با داده‌های گذشته آموزش داده شده و با ورود داده‌های جدید، پیش‌بینی انجام می‌شود. دقت مدل با استفاده از پنجره لغزان پایش می‌شود و در صورت تشخیص رانش مفهوم، به‌روزرسانی مدل صورت می‌گیرد. این یادگیری افزایشی باعث می‌شود مدل تنها در مواقع ضروری به‌روزرسانی شود که ضمن حفظ دقت، نیاز به برچسب‌گذاری کامل داده‌ها کاهش داده می‌شود.

بررسی دو چالش رانش مفهوم و محدودیت برچسب‌گذاری: در این بخش به بررسی دو چالش مهم رانش مفهوم و محدودیت در برچسب‌گذاری داده‌ها پرداخته شده است و در مطالعات انجام‌شده، مسئله داده‌های نامتوازن کلاس‌ها مورد بررسی قرار نگرفته است. غفلت از این جنبه می‌تواند منجر به کاهش دقت مدل‌ها و تضعیف قابلیت تعمیم آن‌ها در جریان‌های داده واقعی شود. در مطالعه [۱۳]، روش $OLIF^3$ را معرفی شده است که با استفاده از الگوریتم KNN و معیارهایی مانند اکثریت همسایه‌ها و چارک سوم، نمونه‌های بدون برچسب غربال می‌شوند و تنها نمونه‌های قابل اعتماد برای آموزش طبقه‌بند اصلی انتخاب می‌شوند تا سازگاری تدریجی مدل با رانش مفهوم حاصل شود و در عین حال نیاز به برچسب‌گذاری کامل کاهش داده شود. در مطالعه [۱۴]، روش $VAE4AS^4$ ارائه شده است که با استفاده از شبکه VAE و پنجره‌های لغزان، داده‌ها بازسازی می‌شوند و داده‌های ناهنجار شناسایی می‌شوند و با به‌کارگیری مکانیزم تشخیص دوگانه، رانش مفهوم کشف شده و مدل بازآموزی می‌شود. در

مفهوم سازگار می‌شود. یادگیری فعال چندمعیاره، راهبردی است که از ترکیب عدم قطعیت، چگالی داده و ارتباط با پنجره لغزان برای انتخاب نمونه‌های بهینه جهت برچسب‌گذاری استفاده کرده و از انتخاب نمونه‌های نویزی جلوگیری می‌کند. تولید داده مصنوعی نوین، روشی است که با ترکیب وزن‌دار نمونه‌های کلاس اقلیت، نویز کنترل‌شده و دنباله آشوبی $Logistic Map$ ، تنوع نمونه‌های تولیدی را افزایش داده و عملکرد مدل را در شرایط عدم توازن و رانش مفهوم بهبود می‌بخشد. برچسب‌گذاری مبتنی بر RBF ، مکانیزمی است که با استفاده از شباهت ساختاری نمونه‌ها، برچسب‌های احتمالاتی تولید کرده و خطای انتشار برچسب را کاهش می‌دهد. تابع هزینه ترکیبی، ترکیبی از $focal loss$ ، وزن‌دهی پویا کلاس‌ها و بهینه‌سازی $Adam$ است که همزمان بر نمونه‌های دشوار، کلاس اقلیت و نمونه‌های نویزی ناشی از رانش مفهوم تمرکز دارد.

اهداف پژوهش حاضر به‌صورت زیر تعریف می‌شوند:

۱- ارائه یک روش نیمه‌نظارتی برای طبقه‌بندی جریان داده با قابلیت سازگاری با رانش مفهوم و تغییرات پویای توزیع داده‌ها.

۲- کاهش وابستگی مدل به داده‌های برچسب‌دار و مدیریت محدودیت در فرآیند برچسب‌گذاری از طریق بهره‌گیری از راهکارهای یادگیری نیمه‌نظارتی و یادگیری فعال.

۳- بهبود عملکرد مدل در مواجهه با داده‌های نامتوازن و افزایش دقت در شناسایی نمونه‌های متعلق به کلاس اقلیت.

۴- افزایش تعمیم‌پذیری و پایداری مدل در محیط‌های واقعی جریان داده و ارائه عملکرد قابل اعتماد در شرایط پویا و متغیر.

ساختار ادامه مقاله به این صورت تنظیم شده است. در بخش دوم، مطالعات و پژوهش‌های پیشین مرور می‌شوند. در بخش سوم، روش پیشنهادی ارائه می‌شود. در بخش چهارم، نتایج حاصل از ارزیابی روش پیشنهادی تحلیل و بررسی می‌شوند. در بخش پنجم، جمع‌بندی و نتیجه‌گیری پژوهش بیان می‌شود. در نهایت، جهت‌گیری‌های پژوهشی آینده ارائه می‌شوند.

۲- کارهای مرتبط

مطالعات انجام‌شده به سه گروه طبقه‌بندی می‌شوند. در گروه اول تنها به چالش رانش مفهوم پرداخته می‌شود. در گروه دوم دو چالش رانش مفهوم و محدودیت برچسب‌گذاری در نظر گرفته می‌شود و در ادامه، در گروه سوم هر سه چالش رانش مفهوم، داده‌های نامتوازن و محدودیت برچسب‌گذاری بررسی می‌شوند.

مطالعه [۱۵]، مدل SSALD^۵ معرفی شده است که بر پایه شبکه عصبی پایه شعاعی با پنجره لغزان و استراتژی پرس‌وجوی مبتنی بر خطای همواری طراحی شده است، به‌گونه‌ای که نمونه‌های بدون برچسب با بیشترین زیان همواری برای برچسب‌گذاری انتخاب می‌شوند و با تنظیم تطبیقی نودها و منظم‌سازی نیمه‌نظارتی، مدل برای مقابله با رانش مفهوم و محدودیت برچسب‌گذاری بهینه‌سازی می‌شود. در مطالعه [۱۶]، چارچوبی معرفی شده است که در آن با برآورد عدم‌اطمینان نمونه‌های بدون برچسب و محاسبه p-value، رانش مفهوم تشخیص داده می‌شود و با افزودن نمونه‌های مطمئن به فرآیند آموزش، مدل به‌روزرسانی می‌شود.

بررسی سه چالش رانش مفهوم، داده‌های نامتوازن و محدودیت برچسب‌گذاری: مطالعاتی که به‌طور همزمان هر سه چالش مطرح‌شده را مورد بررسی قرار دهند، در گذشته اندک بوده‌اند. با این حال، در سال‌های اخیر، پژوهش‌های قابل توجهی به این حوزه معطوف شده‌اند که تلاش می‌شود رویکردهایی جامع و مقاوم در برابر این چالش‌ها ارائه شوند. شایان ذکر است که در مقایسه با روش‌های پیشین، در اکثر مجموعه داده‌ها عملکرد روش پیشنهادی به‌مراتب بهتر نشان داده شده است.

در مطالعه [۱۶]، چارچوب CBAL^۶ با بهره‌گیری از آشکارساز KSPROB برای تشخیص رانش مفهوم و خوشه‌بندی CLUS برای کاهش هزینه برچسب‌گذاری، به همراه نمونه‌برداری افزایشی برای رفع داده‌های نامتوازن کلاس‌ها، طراحی شده است. در مطالعه [۳]، روش SSALD-EIS^۷، با ترکیب یادگیری نیمه‌نظارتی، انتخاب فعال نمونه‌ها و به‌روزرسانی پویا به‌کار گرفته شده است. در هر گام از ورود داده جدید، نمونه‌ها وارد پنجره لغزان می‌شوند و برچسب نمونه‌های بدون برچسب با استفاده از دو منبع پیش‌بینی (خوشه و مدل تجمیعی) تخمین زده می‌شود. سپس یکی از نمونه‌های بدون برچسب که بیشترین ارزش اطلاعاتی را دارد، برای برچسب‌گذاری انتخاب می‌شود. این انتخاب بر پایه عدم قطعیت پیش‌بینی‌ها و اهمیت تعلق به کلاس اقلیت انجام می‌شود تا مسئله داده‌های نامتوازن کاهش داده شود. برای مقابله با رانش مفهوم، مدل تجمیعی به‌صورت پیوسته به‌روزرسانی می‌شود. در مطالعه [۱۷]، روش AACE-DI^۸ با ترکیب یک طبقه‌بند پایدار و طبقه‌بندهای پویا، به‌طور همزمان با سه چالش اصلی جریان داده‌ها یعنی داده‌های نامتوازن، رانش مفهوم و یادگیری نیمه‌نظارتی مقابله می‌شود. برای مقابله با داده‌های نامتوازن داده‌ها، از یک بافر

کشویی استفاده می‌شود که نمونه‌های برچسب‌خورده هر کلاس در آن ذخیره می‌شوند و در صورت کمبود نمونه از کلاس اقلیت، نمونه‌های لازم برای متعادل‌سازی انتخاب می‌شوند. برای مقابله با رانش مفهوم، طبقه‌بندهای پویا بر اساس بلوک‌های جدید داده ساخته می‌شوند و در صورتی که عملکرد طبقه‌بند جدید بهتر باشد، جایگزین طبقه‌بندهای قدیمی می‌شوند تا مدل همواره با تغییرات داده همگام بماند. همچنین در این روش از خوشه‌بندی سلسله‌مراتبی برای شناسایی ساختار داده‌ها استفاده می‌شود و با استراتژی خودکار تنظیم آستانه عدم قطعیت، نمونه‌های مهم و نماینده برای برچسب‌گذاری انتخاب می‌شوند. این انتخاب باعث صرفه‌جویی در منابع برچسب‌گذاری می‌شود. مطالعه [۱۳]، الگوریتم LEPID با هدف مقابله با چالش‌های رانش مفهوم، داده‌های نامتوازن داده‌ها و یادگیری نیمه‌نظارتی طراحی شده است. در این الگوریتم، ابتدا داده‌ها با استفاده از k-means به چند خوشه تقسیم می‌شوند و خلاصه آماری ویژگی‌های هر خوشه در قالب میکروکلاسترها نگه‌داری می‌شود. برای به‌روزرسانی مدل، اگر پیش‌بینی درست باشد، وزن میکروکلاستر بسته به فاصله نمونه از مرکز افزایش داده می‌شود و اگر پیش‌بینی نادرست باشد، وزن کاهش داده شده یا میکروکلاستر شکسته می‌شود. برای برچسب‌گذاری نیز از ترکیب چند استراتژی استفاده می‌شود: انتخاب تصادفی نمونه‌ها برای حفظ توزیع داده، انتخاب نمونه‌ها بر اساس نرخ خطای میکروکلاستر برای شناسایی رانش مفهوم، و استفاده از آنروپی محلی برای یافتن نمونه‌های مهم. در مطالعه [۱۸]، روش DDAL^۹ با استفاده از پنجره زمانی لغزان و خوشه‌بندی K-means، داده‌ها به‌صورت پیوسته پردازش می‌شوند و نمونه‌های کلیدی هر کلاس برای برچسب‌گذاری انتخاب می‌شوند. این نمونه‌ها برچسب‌گذاری شده و به سایر نمونه‌های خوشه گسترش داده می‌شوند تا برای آموزش مدل استفاده شوند. برای مقابله با داده‌های نامتوازن، به نمونه‌های نادر کلاس اقلیت اولویت بیشتری داده می‌شود و برای شناسایی رانش مفهوم از الگوریتم ADWIN بهره گرفته می‌شود.

مطالعات پیشین رویکردهای متنوعی را برای مدیریت همزمان چالش‌های رانش مفهوم، محدودیت برچسب‌گذاری و عدم‌توازن کلاس‌ها در داده‌های جریانی پیشنهاد داده‌اند. در جدول مقایسه مطالعات (۱)، خلاصه‌ای از این روش‌ها و ویژگی‌های کلیدی آن‌ها شامل مزایا، معایب و چالش‌های پوشش داده شده، مورد مقایسه قرار گرفته است.

جدول ۱- جدول مقایسه مطالعات

مرجع	روش	چالش‌های پوشش داده‌شده	مزایا	معایب
[11]	OEC	رانش مفهوم	استفاده از ساختار دولایه و تنظیم پویای وزن طبقه‌بندها، مقاومت مناسب در برابر نویز و رانش مفهوم	عدم بررسی داده‌های نامتوازن و محدودیت برچسب‌گذاری
[12]	مبتنی بر یادگیری گروهی	رانش مفهوم	به‌روزرسانی مدل فقط در زمان نیاز، کاهش هزینه محاسباتی و حفظ دقت	عدم توجه به داده‌های نامتوازن و انتخاب نمونه برای برچسب‌گذاری
[13]	OLIF	رانش مفهوم + محدودیت برچسب	انتخاب نمونه‌های قابل‌اعتماد با استفاده از KNN و کاهش نیاز به برچسب‌گذاری کامل	عدم مدیریت داده‌های نامتوازن
[14]	VAE4AS	رانش مفهوم + محدودیت برچسب	استفاده از VAE و مکانیزم تشخیص دوگانه برای شناسایی رانش مفهوم و داده‌های ناهنجار	عدم بررسی مشکل عدم‌توازن کلاس‌ها
[15]	SSALD	رانش مفهوم + محدودیت برچسب	استفاده از استراتژی پرس‌وجوی تطبیقی و منظم‌سازی نیمه‌نظارتی	عدم توجه به داده‌های نامتوازن
[16]	مبتنی بر p-value	رانش مفهوم + محدودیت برچسب	تشخیص رانش مفهوم با برآورد عدم‌اطمینان و به‌روزرسانی تدریجی مدل	عدم مدیریت عدم‌توازن داده‌ها
[16]	CBAL	رانش مفهوم + داده‌های نامتوازن + محدودیت برچسب	استفاده از خوشه‌بندی، نمونه‌برداری افزایشی و آشکارساز رانش مفهوم	پیچیدگی محاسباتی بالا در خوشه‌بندی و تنظیم پارامترها
[3]	SSALE-EIS	رانش مفهوم + داده‌های نامتوازن + محدودیت برچسب	انتخاب فعال نمونه‌ها بر اساس عدم‌قطعیت و اهمیت کلاس اقلیت، به‌روزرسانی پویا	وابستگی به کیفیت پیش‌بینی اولیه و حساسیت به انتخاب پنجره لغزان
[17]	AACE-DI	رانش مفهوم + داده‌های نامتوازن + محدودیت برچسب	ترکیب طبقه‌بندهای پایدار و پویا، تنظیم خودکار آستانه عدم‌قطعیت	افزایش پیچیدگی مدل و هزینه پردازشی
[13]	LEPID	رانش مفهوم + داده‌های نامتوازن + محدودیت برچسب	استفاده از میکروکلاسترها و چندین استراتژی انتخاب نمونه	حساسیت به تعداد خوشه‌ها و پارامترهای خوشه‌بندی
[18]	DDAL	رانش مفهوم + داده‌های نامتوازن + محدودیت برچسب	استفاده از پنجره لغزان، خوشه‌بندی و ADWIN برای تشخیص رانش مفهوم	وابستگی عملکرد به کیفیت خوشه‌بندی و انتخاب نمونه‌های کلیدی

۳-روش پیشنهادی

الگوریتم (۳-۱): یادگیری فعال برای برچسب‌گذاری

Step	Algorithm
Input	data_window, label_window, label_index, window_size, label_budget, sigma
Output	label_index, labeled_data, labeled_data_y
1	Calculate labeling limit: num = window_size × label_budget
2	Calculate labeling probability: label_probability ← 1 - sum(label_index) / num
3	Calculate uncertainty: margin[i] ← pre[i] - 0.5 uncertainty[i] ← -margin[i]
4	Calculate density via Eq. (3-5)
5	Calculate selection_score: selection_score[i] ← uncertainty_weight × uncertainty[i] + (1 - uncertainty_weight) × density[i]
6	Ignore labeled samples: if label_index[i] == 1 THEN selection_score[i] = -∞
7	Select top 5 samples with highest score.
8	Randomly pick one from them.
9	Mark it as labeled: Label_index[chosen] = 1
10	Update label_index, labeled_data, labeled_data_y

با توجه به الگوریتم‌های (۳-۱) یادگیری فعال برای برچسب‌گذاری، (۳-۲) تنظیم تطبیقی گره و مقابله با رانش مفهومی، و (۳-۳) مقابله با عدم تعادل داده‌ها، چارچوب روش پیشنهادی به گونه‌ای تدوین شده است که پیاده‌سازی و ارائه روش پیشنهادی را به شکلی شفاف و قابل فهم تضمین می‌کند. این رویکرد جامع، تمامی جنبه‌های تحقیق را پوشش داده و به درک عمیق‌تر روش کمک شایانی می‌نماید.

در روش پیشنهادی، ویژگی‌های ورودی پس از تبدیل به مقادیر عددی، از طریق جایگزینی مقادیر نامعتبر با مقدار صفر پس از تخصیص موقت پاک‌سازی شدند. برچسب‌ها نیز با استفاده از کدگذاری دودویی بازنمایی شدند. مدل RBF ابتدا با استفاده از ۲۰۰ نمونه اولیه مقداردهی شده و در ادامه، مدل به صورت جریانی بقیه داده‌ها را پردازش کرده و پیش‌بینی‌های لازم را انجام خواهد داد. علاوه بر این، پارامترهای مدل نیز تعیین گردید؛ به این صورت که مقدار اولیه برای α برابر با ۰/۴ در نظر گرفته شد و پارامتر σ در بازه 2^{-8} تا 2^8 مورد ارزیابی قرار گرفت.

تابع خطای focal که علاوه بر تمرکز بر نمونه‌های سخت، توزیع نامتوازن داده‌ها را نیز اصلاح می‌کند. شکل (۱) فلوچارت روش پیشنهادی را نشان می‌دهد.

۳-۱- پنجره لغزان و بودجه برچسب گذاری

پنجره لغزان در این مدل به کار گرفته می‌شود تا آخرین نمونه‌های داده برای آموزش مدل نگهداری شوند. اندازه آن تعیین می‌کند که چه تعداد از جدیدترین نمونه‌ها در حافظه باقی گذاشته شوند و داده‌های قدیمی‌تر حذف گردند. در هر به‌روزرسانی، ابتدا نمونه‌های قدیمی حذف و سپس داده‌های جدید به انتهای پنجره افزوده می‌شوند تا مدل همواره با داده‌های اخیر تطبیق داده شود. بودجه برچسب‌گذاری، که با B نشان داده می‌شود، میزان داده‌هایی را که می‌توانند برچسب واقعی دریافت کنند مشخص می‌کند. این پارامتر برای کاهش هزینه و زمان برچسب‌گذاری و استفاده مؤثر از داده‌ها در یادگیری نیمه‌نظارتی تعریف شده است. در هر لحظه با ورود داده‌های جدید، تصمیم گرفته می‌شود که آیا نمونه‌ای باید برچسب‌گذاری شود یا خیر، که این تصمیم به‌صورت احتمالی و مطابق با رابطه (۱) اتخاذ می‌گردد. در این رابطه، n_i تعداد نمونه‌هایی است که در حال حاضر در پنجره W_s برچسب‌خورده‌اند، BN حداکثر تعداد مجاز نمونه‌های برچسب‌خورده در پنجره است و P احتمال برچسب‌گذاری نمونه جدید را نشان می‌دهد.

$$P = 1 - \frac{n_i}{BN} \quad (1)$$

۳-۲- یادگیری فعال برای برچسب گذاری

در صورتی که بودجه برچسب‌گذاری کافی در دسترس باشد، نمونه‌های مناسب برای برچسب‌گذاری انتخاب می‌شوند. در این فرآیند، ابتدا خروجی مدل برای تمامی نمونه‌های موجود در پنجره محاسبه می‌گردد. در مسائل دودویی، مقدار 0.5 به‌عنوان مرز تصمیم‌گیری در نظر گرفته می‌شود. نمونه‌هایی که احتمال پیش‌بینی آن‌ها به 0.5 نزدیک‌تر است، بیشترین میزان عدم قطعیت را دارند. معیار حاشیه برای هر نمونه x_i مطابق معادله (۲) تعریف می‌شود، که در آن p_i نشان‌دهنده احتمال پیش‌بینی‌شده مدل برای تعلق نمونه x_i به یک کلاس است. هرچه مقدار حاشیه بزرگ‌تر باشد، اطمینان مدل از پیش‌بینی خود بیشتر تلقی می‌شود. به‌منظور افزایش احتمال انتخاب نمونه‌های مبهم برای برچسب‌گذاری، معیار عدم قطعیت طبق معادله (۳) تعریف می‌گردد؛ به این ترتیب، نمونه‌هایی با مقادیر بالاتر معیار عدم قطعیت به‌عنوان گزینه‌های مناسب‌تر برای برچسب‌گذاری

الگوریتم (۲-۳): تنظیم تطبیقی گره و مقابله با رانش مفهومی

Step	Algorithm
Input	data_window, centers, window_size, sigma, num_centers
Output	Update centers
1	Calculate representativeness with Eq. (3-11)
2	Find center with minimum representativeness: $\min_index = \text{argmin}(\text{representative})$
3	While $\text{representative}[\min_index] < 0.5$ and $\text{round} < \text{num_centers}$:
4	For each data point i in data_window:
5	Compute kernel similarity with all centers: $\text{rep_ci} = \text{kernel_fun}(\text{data_window}[i], \text{centers}, \sigma)$ $\text{representative}[i] = 1 - \max(\text{rep_ci})$
6	Compute kernel similarity with all data points: $\text{eff_ci} = \text{kernel_fun}(\text{data_window}[i], \text{data_window}, \sigma)$ Average similarity: $\text{eff_ci} = \text{mean}(\text{eff_ci})$ Add to representativeness: $\text{representative}[i] += \text{eff_ci}$
7	Find data point with maximum representativeness: $\max_index = \text{argmax}(\text{representative})$
8	Replace least representative center with that data point: $\text{centers}[\min_index] = \text{data_window}[\max_index]$
9	Update centers

الگوریتم (۳-۳): مقابله با عدم تعادل

Step	Algorithm
Input	labeled_data, labeled_data_y, labeled_weight
Output	Update labeled_data, labeled_data_y, updated weight
1	Separate samples into minority and majority classes based on labels.
2	Check imbalance: If the minority class has fewer samples than the majority class: $N_{\text{synthetic}} = \text{Majority} - \text{Minority} $
3	Generate synthetic samples: $x_{\text{synthetic}} = w_0x_0 + w_1x_1 + w_2x_2 + \epsilon$
4	Update labeled_data, labeled_data_y, labeled_weight

ساختار شبکه عصبی RBFNN شامل سه لایه اصلی ورودی، پنهان و خروجی در نظر گرفته شده است. لایه پنهان دارای یک لایه خطی، نرمال‌سازی لایه‌ای، تابع فعال‌سازی ReLU و Dropout است. در لایه اول، ۲۵۶ نورون با نرخ Dropout برابر 0.3 ، در لایه دوم ۱۲۸ نورون با نرخ 0.2 و در لایه سوم ۶۴ نورون در نظر گرفته شده است. در انتها، احتمال تعلق هر نمونه به کلاس مربوطه توسط لایه خروجی با تابع فعال‌سازی سیگموید در بازه $[0, 1]$ محاسبه می‌شود. این روش شامل چهار بخش اصلی است: یادگیری فعال برای برچسب‌گذاری داده‌ها، تنظیم تطبیقی گره‌ها و مقابله با رانش مفهوم، تولید نمونه‌های مصنوعی با استفاده از ترکیب وزن‌دار نمونه‌های کلاس اقلیت به همراه نویز و استفاده از

فعال در نظر گرفته می‌شوند.

$$\text{margin}(x_i) = |\hat{p}_i - 0.5| \quad (2)$$

$$\text{Uncertainty}(x_i) = -\text{margin}(x_i) \quad (3)$$

مدل برای یک نمونه بدون برچسب برابر با مجموع وزن‌دار برچسب‌های نمونه‌هایی است که در اطراف آن نمونه قرار دارند. در این مدل، داده‌ها در یک پنجره لغزان نگهداری شده و برخی از نمونه‌ها برچسب ندارند. برچسب شباهت برای نمونه‌های بدون برچسب محاسبه می‌شود تا امکان آموزش مدل نیمه‌نظارتی فراهم گردد. برای هر نمونه‌ی بدون برچسب در پنجره‌ی لغزان، طبق معادله (۶) یک برچسب هموار یا تقریبی به دست می‌آید، که در آن y_j برچسب واقعی نمونه‌ی x_j در پنجره‌ی برچسب را نشان می‌دهد S_{ij} میزان شباهت بین نمونه‌ی بدون برچسب x_i و نمونه‌ی برچسب‌خورده‌ی x_j محاسبه می‌گردد w_{ij} وزن نمونه است که میزان تازگی آن نمونه را مشخص می‌کند و S_i مقدار برچسب هموار برای نمونه‌ی بدون برچسب x_i در نظر گرفته می‌شود. برای تعیین میزان شباهت دو نمونه، مطابق معادله (۷) از تابعی موسوم به هسته‌ی پایه‌ی شعاعی استفاده می‌گردد. این تابع عددی در بازه‌ی صفر تا یک تولید می‌کند که بیانگر میزان شباهت دو نمونه است و هرچه نمونه‌ها به یکدیگر نزدیک‌تر باشند، مقدار خروجی تابع به یک نزدیک‌تر تلقی می‌شود. در این معادله، فاصله اقلیدسی بین دو نمونه محاسبه شده و S_{ij} میزان شباهت بین نمونه‌های x_i و x_j را نشان می‌دهد و σ عرض توابع پایه شعاعی در نظر گرفته می‌شود.

$$S_i = \frac{\sum_{j \in W_i} w_j S_{ij} y_j}{\sum_{j \in W_i} w_j S_{ij}} \quad (6)$$

$$S_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right) \quad (7)$$

۳-۴- تنظیم تطبیقی گره و مقابله با رانش مفهوم

برای مقداردهی اولیه مراکز در یک مدل شبکه‌ای مانند RBFNN از روشی استفاده می‌شود که نسخه‌ای پیشرفته از K-means محسوب می‌شود. در روش K-means++، علاوه بر احتمال انتخاب مبتنی بر فاصله، معیار تنوع میان مراکز نیز در نظر گرفته می‌شود. فرآیند مقداردهی اولیه به این صورت انجام می‌گیرد که ابتدا یک نمونه به‌صورت تصادفی از میان داده‌ها انتخاب شده و به‌عنوان نخستین مرکز در نظر گرفته می‌شود. سپس برای هر نمونه فاصله‌ی آن تا نزدیک‌ترین مرکز انتخاب‌شده توسط معادله (۸) محاسبه می‌گردد، که در آن C مجموعه‌ی مراکز انتخاب‌شده تا آن لحظه است. بر اساس این مقادیر، یک توزیع احتمال برای انتخاب مرکز جدید طبق معادله (۹) تعریف می‌شود، که برابر فاصله آن نمونه تقسیم بر مجموع فاصله‌هاست؛ هر مقدار این بردار احتمال

با این حال، استفاده صرف از معیار عدم قطعیت ممکن است منجر به انتخاب داده‌های نویزی یا دورافتاده شود. برای جلوگیری از این مشکل، از معیار چگالی نیز در کنار آن بهره گرفته می‌شود. نمونه‌هایی که در نواحی متراکم‌تر داده قرار دارند، نماینده بهتری از توزیع کلی داده تلقی شده و برچسب‌گذاری آن‌ها اطلاعات مفیدتری برای مدل فراهم می‌کند. چگالی هر نمونه نشان داده می‌شود تا میزان نزدیکی سایر نمونه‌ها به آن مشخص شود؛ هرچه چگالی بیشتر باشد، اهمیت آن نمونه برای یادگیری مدل بالاتر در نظر گرفته می‌شود. مقدار چگالی طبق معادله (۴) محاسبه می‌گردد؛ در این معادله N تعداد نمونه‌های موجود در پنجره داده جاری است. صورت کسر فاصله اقلیدسی بین نمونه‌های x_i و x_j را نشان می‌دهد و پارامتر σ بر اساس فاصله مراکز محاسبه می‌گردد. در نهایت، معیار انتخاب نمونه برای برچسب‌گذاری طبق معادله (۵) به‌صورت ترکیبی از دو عامل عدم قطعیت و چگالی تعریف می‌شود. طبق این معادله پارامتر α به‌عنوان ضریب متعادل‌کننده بین معیار عدم قطعیت و چگالی داده عمل می‌کند و میزان تأثیر هر یک از این دو مؤلفه را در فرآیند انتخاب نمونه تعیین می‌نماید. مقادیر بزرگ‌تر α باعث می‌شود الگوریتم تمرکز بیشتری بر نمونه‌های دارای عدم قطعیت بالا داشته باشد، در حالی که مقادیر کوچک‌تر آن موجب می‌شود نمونه‌های واقع‌شده در نواحی پرتراکم داده اولویت بیشتری برای برچسب‌گذاری پیدا کنند. نمونه‌ای که بالاترین امتیاز را کسب کند، برای برچسب‌گذاری انتخاب شده و پس از دریافت برچسب واقعی، به مجموعه آموزشی افزوده می‌گردد.

$$\text{Density}(x_i) = \exp\left(-\frac{\frac{1}{N} \sum_{j=1}^N \|x_i - x_j\|^2}{\sigma}\right) \quad (4)$$

$$\text{SelectionScore}(x_i) = \alpha \cdot \text{Uncertainty}(x_i) + (1 - \alpha) \cdot \text{Density}(x_i) \quad (5)$$

۳-۳- محاسبه برچسب هموار

فرضیه‌ی همواری یکی از مفاهیم رایج در یادگیری نیمه‌نظارتی تلقی می‌شود. براساس این فرضیه، خروجی مدل برای نمونه‌های نزدیک به هم مشابه یا یکسان در نظر گرفته می‌شود. پیش‌بینی

عرض توابع پایه شعاعی در نظر گرفته می‌شود. معادله (۱۱) از دو بخش تشکیل شده است: بخش اول میزان شباهت مرکز c_i با مراکز دیگر را مشخص می‌کند؛ هرچه شباهت بیشتر باشد، نمایندگی کمتر در نظر گرفته می‌شود. بخش دوم میزان نزدیکی c_i به نمونه‌های موجود در پنجره را می‌سنجد؛ هرچه نزدیکی بیشتر باشد، نمایندگی بالاتر تلقی می‌شود. اگر مرکزی نمایندگی پایینی داشته باشد، با نمونه‌ای جایگزین می‌شود که نمایندگی بالاتری دارد. معادله (۱۲) نشان می‌دهد که مرکز جدید c_i همان نمونه‌ای است که بیشترین نمایندگی را دارد.

$$R(c_i) = 1 - \max_{j \in C, j \neq i} \exp\left(-\frac{\|c_i - c_j\|^2}{2\sigma^2}\right) + \quad (11)$$

$$\sum_{j \in W_s} \exp\left(-\frac{\|c_i - x_j\|^2}{2\sigma^2}\right)$$

$$c_i = \arg \max_{j \in W_s} R(x_j) \quad (12)$$

در شبکه‌های عصبی با تابع پایه شعاعی، هر نورون پنهان دارای یک تابع پایه شعاعی است که معمولاً به شکل تابع گاوسی تعریف می‌شود. این تابع تعیین می‌کند که هر مرکز نورون چه میزان تأثیری بر نمونه‌های داده داشته باشد. یکی از پارامترهای مهم این تابع، عرض هسته σ است که برای مشخص کردن میزان گستردگی یا تمرکز تابع گاوسی به کار گرفته می‌شود. در صورتی که عرض هسته بسیار کوچک انتخاب شود، تابع گاوسی باریک شده و تنها نمونه‌هایی که فاصله بسیار کمی با مرکز نورون دارند، تأثیر قابل توجهی دریافت می‌کنند. این وضعیت باعث افزایش حساسیت مدل می‌شود. از سوی دیگر، اگر عرض هسته بسیار بزرگ در نظر گرفته شود، تابع گاوسی گسترده و تخت می‌گردد. در این حالت، نورون‌ها تأثیر خود را بر بخش وسیعی از فضای داده‌ها اعمال می‌کنند و مدل ممکن است توانایی تشخیص جزئیات دقیق داده‌ها را از دست بدهد که این امر به بیش‌تعمیم منجر می‌شود. در این روش، مقدار σ بر اساس معادله (۱۳) محاسبه می‌شود، که در آن صورت کسر بزرگترین فاصله اقلیدسی بین دو مرکز نورون c_i و c_j است و H تعداد کل مراکز نورون‌ها (نورون‌های لایه پنهان) را نشان می‌دهد.

$$\sigma = \frac{\max_{i,j \in C} \|c_i - c_j\|}{\sqrt{2H}} \quad (13)$$

۳-۵- مقابله با داده‌های نامتوازن با تولید نمونه تصادفی توسط دنباله آشوب

یکی از چالش‌های مهم در یادگیری ماشین، داده‌های نامتوازن

بیانگر شانس انتخاب آن نمونه به عنوان کاندید مرکز جدید در نظر گرفته می‌شود. نمونه‌هایی که فاصله بیشتری از مراکز فعلی دارند، احتمال بالاتری برای انتخاب شدن به عنوان مرکز بعدی خواهند داشت. به جای انتخاب مستقیم یک مرکز از این توزیع، چندین نامزد به صورت احتمالی انتخاب می‌شوند و برای هر نامزد معیار تنوع محاسبه می‌گردد. برای سنجش تنوع، ابتدا فاصله اقلیدسی بین همه جفت‌های مراکز محاسبه شده، سپس کوچک‌ترین فاصله باقی‌مانده انتخاب می‌گردد، میانگین همه فاصله‌های غیر صفر بین مراکز محاسبه شده و نسبت کوچک‌ترین فاصله به میانگین فاصله‌ها طبق معادله (۱۰) به عنوان نمره تنوع بازگردانده می‌شود. در نهایت، کاندیدی که بیشترین تنوع را ایجاد کند به عنوان مرکز انتخاب می‌گردد. در نتیجه، این روش باعث می‌شود مراکز خوشه‌ها به گونه‌ای انتخاب شوند که هم پوشش خوبی از داده‌ها فراهم گردد و هم بیشترین تنوع بین مراکز ایجاد شود تا خوشه‌ها به هم شبیه نباشند.

$$D(x_j) = \min_{(c \in C)} \|x_j - c\|^2 \quad (8)$$

$$P(x_j) = \frac{D(x_j)}{\sum_{l=1}^N D(x_l)} \quad (9)$$

$$Diversity\ Score = \frac{\min_{i \neq j} d(c_i, c_j)}{\max_{i \neq j} d(c_i, c_j)} \quad (10)$$

شبکه‌های عصبی با تابع پایه شعاعی معمولاً از معیاری مبتنی بر فاصله برای تعیین میزان فعال‌سازی نورون‌های لایه پنهان استفاده می‌شوند. موقعیت مراکز RBF نقش تعیین‌کننده‌ای در دقت پیش‌بینی مدل ایفا می‌کند، زیرا نمونه‌هایی که فاصله زیادی از این مراکز داشته باشند، فعال‌سازی کمی در لایه پنهان دریافت می‌کنند و این موضوع می‌تواند باعث کاهش دقت پیش‌بینی شود. با این حال، هنگام مواجهه با رانش مفهوم، ممکن است مراکز RBF انتخاب‌شده در گذشته با توزیع فعلی داده‌ها همخوانی نداشته باشند. برای مقابله با این چالش، روش‌های تطبیقی پیشنهاد شده‌اند که موقعیت نورون‌ها به روزرسانی می‌شوند تا توزیع جدید داده‌ها شناسایی گردد. این کار با استفاده از معیار مشخص‌شده در معادله (۱۱) انجام می‌گیرد تا میزان ارتباط و نمایندگی هر نورون نسبت به توزیع جاری داده‌ها سنجیده شود، که در آن $R(c_i)$ میزان نمایندگی مرکز c_i را نشان می‌دهد، C مجموعه تمام مراکز RBF در شبکه است، $\|c_i - c_j\|^2$ فاصله اقلیدسی بین مراکز نورون c_i و سایر مراکز c_j را نشان می‌دهد و $\|c_i - x_j\|^2$ فاصله بین مرکز نورون c_i و نمونه داده جدید x_j در پنجره W_s را مشخص می‌کند σ .

$$N_{synthetic} = |S_{majority} - S_{minority}| \quad (۱۴)$$

$$X_{synthetic} = w_1 x_{i1} + w_2 x_{i2} + w_3 x_{i3} + \varepsilon \quad (۱۵)$$

۳-۶- تابع خطا

تابع خطا به عنوان معیاری مورد استفاده قرار می گیرد که نشان دهد پیش بینی شبکه تا چه حد با داده های واقعی فاصله دارد و به اصلاح وزن ها در طول آموزش کمک می کند. ابتدا وزن نمونه های مثبت و منفی متناسب با تعدادشان تنظیم می شود. اگر نمونه های مثبت کم باشند، وزن مثبت طبق معادله (۱۶) افزایش داده می شود تا مدل بر آن ها تمرکز بیشتری داشته باشد. در صورت کم بودن نمونه های منفی، وزن منفی طبق معادله (۱۷) افزایش می یابد تا تعادل بین کلاس ها برقرار گردد. وزن ها سپس به صورت تدریجی طبق معادله (۱۸) به روزرسانی می شوند تا بخشی از وزن قبلی حفظ شود.

جدول ۲- توابع آشوبی

نام	تابع
Logistic Map	$x_{n+1} = r x_n (1 - x_n)$
ICMIC Map	$x_{n+1} = \sin\left(\frac{a}{x_n}\right)$
Sinus Map	$x_{n+1} = 2.3(x_n)^{2\sin(\pi x_n)}$
Sinusoidal Map	$x_{n+1} = a x_n^2 \sin(\pi x_n)$
Tent Map	$x_{n+1} = \begin{cases} \frac{x_n}{0.7} & x_n < 0.7, \\ \frac{10}{3}(1-x_n) & x_n \geq 0.7 \end{cases}$
Circle Map	$x_n = \text{mod}\left(x_n + b + \left(\frac{a}{2\pi}\right) \cdot \sin(2\pi x_k), 1\right)$ $a = 0.5, b = 0.2$
Gauss/ Mouse Map	$x_{n+1} = \begin{cases} 1 & x_n = 0 \\ \frac{1}{\text{mod}(x_n, 1)} & \text{otherwise} \end{cases}$
Piecewise Map	$x_{n+1} = \begin{cases} \frac{x_n}{p}, & 0 \leq x_n < p \\ \frac{x_n - p}{0.5 - p}, & p \leq x_n < 0.5 \\ \frac{1 - p - x_n}{0.5 - p}, & 0.5 \leq x_n < 1 - p \\ \frac{1 - x_n}{p}, & 1 - p \leq x_n < 1 \end{cases}, p = 0.4$

کلاس ها تلقی می شود؛ یعنی زمانی که تعداد نمونه های کلاس اقلیت نسبت به کلاس اکثریت بسیار کمتر در نظر گرفته می شود. این موضوع می تواند موجب کاهش عملکرد مدل، به ویژه در شناسایی کلاس اقلیت گردد. برای مقابله با این مشکل، در این پژوهش یک روش تولید نمونه های مصنوعی بهبود یافته طراحی شده است که با هدف متعادل سازی پنجره داده به کار گرفته می شود. روش های رایج متوازن سازی مانند SMOTE برای داده های ایستا طراحی شده اند و در محیط های پویا که با رانش مفهوم مواجه اند، کارایی کمی دارند. روش پیشنهادی یک روش تولید نمونه های مصنوعی مبتنی بر دنباله آشوبی Logistic Map پیشنهاد کرده است. در این روش به جای استفاده از اعداد تصادفی معمولی، از دنباله های آشوبی استفاده می شود. این دنباله ها به دلیل ویژگی هایی نظیر رفتار غیرخطی، حساسیت به شرایط اولیه و پراکندگی یکنواخت، تنوع بسیار بیشتری نسبت به روش های سنتی ایجاد می کنند. که کاهش چشمگیر تولید داده های تکراری (افزایش فاصله اقلیدسی میان نمونه ها)، حفظ بهتر ساختار آماری کلاس اقلیت و بهبود شناسایی مرز تصمیم در مدل، انعطاف پذیری و پایداری بالاتر در مواجهه با داده های پویا و رانش مفهوم را به دنبال دارند. ابتدا اختلاف تعداد نمونه های کلاس اکثریت و اقلیت طبق معادله (۱۴) محاسبه می گردد، که در آن $N_{synthetic}$ تعداد نمونه های مصنوعی لازم برای ایجاد تعادل را نشان می دهد، $S_{majority}$ مجموعه نمونه های کلاس اکثریت و $S_{minority}$ مجموعه نمونه های کلاس اقلیت در نظر گرفته می شوند. اگر تعداد نمونه های اقلیت کمتر از نمونه های اکثریت باشد، فرآیند تولید نمونه مصنوعی طبق معادله (۱۵) آغاز می گردد. نمونه های مصنوعی با استفاده از ترکیب وزن دار سه نمونه واقعی از کلاس اقلیت به همراه مقدار تصادفی تولید می شوند. به منظور تولید مقادیر شبه تصادفی با کیفیت بالا، از توابع آشوبی جدول (۲) بهره گرفته می شود. برای تولید دنباله ای از مقادیر شبه تصادفی، در اینجا تنها از تابع آشوبی Logistic Map استفاده شده است. ابتدا یک مقدار اولیه به صورت تصادفی از بازه (۰،۱) انتخاب می شود. سپس با به کارگیری توابع آشوبی، دنباله ای از مقادیر آشوبی به طور تکرارشونده تولید می گردد، به طوری که هر مقدار خروجی به عنوان ورودی مرحله ی بعد در نظر گرفته می شود. برای حذف اثرات اولیه، تعدادی از مقادیر ابتدایی دنباله کنار گذاشته می شوند. در ادامه، مقادیر باقی مانده به عنوان مقادیر شبه تصادفی قابل اعتماد در مراحل بعدی تحقیق مورد استفاده قرار می گیرند. توابع آشوبی در جدول (۲) ارائه شده اند.

۴- ارزیابی

این بخش به ارزیابی روش پیشنهادی در مقایسه با روش‌های دیگر پرداخته شده است.

۴-۱- مجموعه داده

در آزمایش‌های انجام شده، سه مجموعه داده واقعی و یک مجموعه داده مصنوعی به کار گرفته شده‌اند. مجموعه داده Covtype یک مجموعه داده بزرگ است که برای پیش‌بینی نوع پوشش جنگلی طراحی شده است. مجموعه داده Electricity یک مجموعه داده پرکاربرد برای پیش‌بینی تغییرات قیمت برق در استرالیا در نظر گرفته شده است. مجموعه داده Drebin یکی از مجموعه‌داده‌های مهم در حوزه پژوهش‌های مرتبط با تشخیص بدافزارهای اندروید محسوب می‌شود. مجموعه داده SINE یک مجموعه داده مصنوعی است که به‌طور گسترده در پژوهش‌های حوزه یادگیری ماشین مورد استفاده قرار می‌گیرد. در جدول (۳)، مجموعه داده‌ها همراه با صفات آن‌ها از جمله تعداد ویژگی‌ها و تعداد نمونه‌ها ارائه شده است.

جدول ۳- مشخصات مجموعه‌داده‌های واقعی و مصنوعی

نوع	مجموعه داده	تعداد ویژگی‌ها	تعداد نمونه‌ها	اندازه کلاس-ها	نسبت کلاس-ها
واقعی	covtype	۵۴	۴۵۲۴۷	(۳۵۷۵۴) (۹۴۹۳)	۳/۷۷
واقعی	electricity	۶	۴۵۳۱۲	(۲۶۰۷۵) (۱۹۲۳۷)	۱/۳۶
واقعی	derbin	۳۰۰	۱۲۹۰۱۳	(۱۲۳۴۵۳) (۵۵۶۰)	۲۲/۲۳
مصنوعی	sine	۴	۱۰۰۰۰	(۵۰۰۱) (۴۹۹۹)	۱

۴-۲- نتایج

در این پژوهش، برای ارزیابی دقیق عملکرد مدل توسعه‌یافته، از مجموعه‌ای از معیارهای سنجش کیفیت بهره گرفته شده است. این معیارها شامل Accuracy، G-Mean، F1-score، Sensitivity و Specificity هستند که هر کدام جنبه‌ای از توانایی مدل در پیش‌بینی صحیح نمونه‌ها را منعکس می‌کنند. یکی از این معیارها، Accuracy است. با وجود اینکه Accuracy به دلیل سادگی و قابل فهم بودن، اغلب به عنوان یک معیار پایه مورد استفاده قرار می‌گیرد، اما در مواجهه با مجموعه داده‌هایی که توزیع کلاس‌های آن‌ها نامتوازن است (یعنی تعداد نمونه‌ها در کلاس‌های مختلف به

$$W_{pos} = \frac{1}{2 \times pos_ratio + \varepsilon} \quad (16)$$

$$W_{neg} = \frac{1}{2 \times (1 - pos_ratio) + \varepsilon} \quad (17)$$

$$\begin{aligned} new\ W_{pos} &= \partial * W_{pos} + (1 - \partial) * new\ W_{pos} \\ new\ W_{neg} &= \partial * W_{neg} + (1 - \partial) * new\ W_{neg} \end{aligned} \quad (18)$$

پیش از محاسبه خطا، وزن‌ها با توجه به توزیع جدید داده‌ها به‌روز می‌شوند و در محاسبه focal loss ضرب می‌گردند تا کلاس‌های کم‌تعداد اهمیت بیشتری پیدا کنند. در این مرحله، احتمال پیش‌بینی صحیح مدل برای هر نمونه توسط معادله (۱۹) محاسبه می‌شود و سپس focal_weight طبق معادله (۲۰) به‌دست آورده می‌شود.

$$p_t = y_{true} \cdot y_{pred} + (1 - y_{true}) \cdot (1 - y_{pred}) \quad (19)$$

$$focal_weight = (1 - p_t^{(i)})^\gamma \quad (20)$$

در نهایت focal loss طبق معادله (۲۱) با وزن‌دهی پویا تشکیل می‌شود که در واقع ترکیبی از دو بخش است: بخش اول focal_weight و بخش دوم عبارت داخل کروشه که در ادامه توضیح داده می‌شود. اگر مدل مطمئن باشد، وزن توسط focal_weight کاهش می‌شود و اگر پیش‌بینی اشتباه باشد، وزن افزایش می‌یابد. عبارت داخل کروشه ترکیبی از دو بخش در نظر گرفته می‌شود. در یادگیری دوکلاسه، هر نمونه مثبت یا منفی تلقی می‌شود. بخش اول آن برای نمونه‌های مثبت و بخش دوم برای نمونه‌های منفی در نظر گرفته شده و هرکدام در وزن‌های مربوط به خود W_{pos} و W_{neg} ضرب می‌شوند. با این وزن‌دهی، اهمیت هر نمونه توسط مدل تنظیم می‌گردد.

$$L = -\frac{1}{N} \sum_{i=1}^N \left(1 - p_t^{(i)} \right)^\gamma \left[y_i \cdot \log(\hat{y}_i + \varepsilon) \cdot w_i \cdot W_{pos} + (1 - y_i) \cdot \log(1 - \hat{y}_i + \varepsilon) \cdot w_i \cdot W_{neg} \right] \quad (21)$$

در نهایت، این روش برای به‌روزرسانی وزن‌های خروجی از الگوریتم Adam استفاده می‌شود. برای هر وزن، نرخ یادگیری خاص خود توسط Adam تنظیم می‌گردد. آموزش تا زمانی ادامه می‌یابد که یکی از شرایط زیر برقرار شود: تغییر بسیار کم در تابع خطا مشاهده شود یا به تعداد حداکثری دوره‌های آموزش دست یافته شود.

$$G-Mean = \sqrt{Sensitivity \times Specificity} \quad (27)$$

همان‌طور که در جدول (۴) مشاهده می‌شود، عملکرد روش پیشنهادی در مقایسه با روش‌های SSALD، AACE-DI، LEPID و SSALE-EIS بر روی چهار مجموعه داده Covetype، Electricity، Drebin و Sine با استفاده از معیارهای Accuracy، Sensitivity و Specificity ارزیابی شده است. انتخاب این معیارها به دلیل حساسیت آن‌ها نسبت به عدم‌توازن کلاس‌ها و تغییرات توزیع داده در جریان‌های داده‌ای انجام شده است؛ به‌ویژه G-Mean و F1-score که عملکرد مدل را در سطح متوازن بین کلاس‌ها منعکس می‌کنند و شاخص مناسبی برای ارزیابی مدل‌های یادگیری در شرایط داده‌های نامتوازن محسوب می‌شوند.

برخلاف روش‌های موجود LEPID که داده‌های نامتوازن را نادیده گرفته یا سطحی مدیریت می‌کنند، رویکرد پیشنهادی با ترکیب عدم‌قطعیت و چگالی کیفیت انتخاب نمونه‌ها را در یادگیری فعال به‌طور چشمگیری بهبود داده و از انتخاب نمونه‌های نویزی

شدت متفاوت است)، می‌تواند تصویری همراه‌کننده از عملکرد واقعی مدل ارائه دهد. به همین دلیل، برای غلبه بر محدودیت‌های Accuracy در داده‌های نامتوازن، در این پژوهش، علاوه بر محاسبه این معیار، تمرکز ویژه‌ای بر استفاده از معیارهایی شده است که به طور خاص نسبت به کلاس اقلیت (کلاسی که تعداد نمونه‌های آن کمتر است) حساسیت بیشتری دارند. معیارهای ذکر شده در بالا، با استفاده از معادلات زیر قابل محاسبه هستند.

$$Accuracy = \frac{TN + TP}{TN + TP + FN + FP} \quad (22)$$

$$Recall = \frac{TP}{TP + FN} \quad (23)$$

$$Sensitivity = \frac{TP}{TP + FN} \quad (24)$$

$$Specificity = \frac{TN}{TN + FP} \quad (25)$$

$$F1_Score = \frac{2 * (Recall * Precision)}{Recall + Precision} \quad (26)$$

جدول ۴- نتایج معیارهای ارزیابی

LEPID [۱۳]	AACE-DI [۱۷]	SSALE-EIS [۳]	SSALD [۱۵]	روش پیشنهادی	معیار ارزیابی	مجموعه داده
-	-	-	۰/۹۳۸	۰/۹۵۵	Accuracy	Covetype
-	-	-	۰/۹۶۲	۰/۹۷۴	F1-score	
۰/۸۶۳	۰/۵۶۶	۰/۷۲۲	۰/۹۶۷	۰/۹۷۰	G-mean	
۰/۸۷۸	۰/۲۱۳	۰/۶۷۷	۰/۹۵۵	۰/۹۸۱	Sensitivity	
-	-	-	۰/۹۷۰	۰/۹۷۲	Specificity	
-	-	-	۰/۹۵۴	۰/۹۸۴	Accuracy	Derbin
-	-	-	۰/۹۴۱	۰/۹۷۸	F1-score	
-	-	-	۰/۹۶۴	۰/۹۸۷	G-mean	
-	-	-	۱	۱	Sensitivity	
-	-	-	۰/۹۲۸	۰/۹۷۴	Specificity	
-	-	-	۰/۸۲۹	۰/۸۹۳	Accuracy	Sine
-	-	-	۰/۸۲۶	۰/۸۹۴	F1-score	
-	-	-	۰/۸۲۹	۰/۸۹۳	G-mean	
-	-	-	۰/۸۱۵	۰/۸۹۷	Sensitivity	
-	-	-	۰/۸۴۲	۰/۸۹۰	Specificity	
-	-	-	۰/۷۹۴	۰/۷۷۲	Accuracy	Electricity
-	-	-	۰/۷۷۰	۰/۷۹۸	F1-score	
۰/۶۹۴	۰/۷۵۲	۰/۸۳۶	۰/۷۸۴	۰/۸۰۷	G-mean	
۰/۷۱۱	۰/۵۳۳	۰/۸۵۵	۰/۷۵۹	۰/۷۸۸	Sensitivity	
-	-	-	۰/۷۹۳	۰/۷۹۶	Specificity	

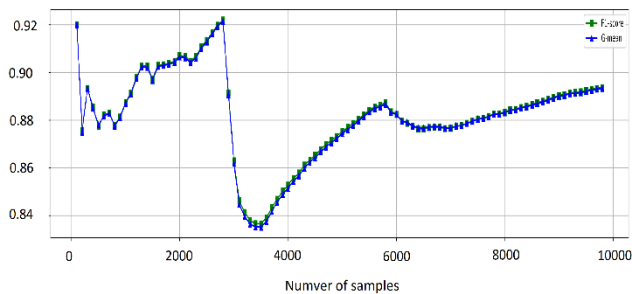
جلوگیری می‌کند. همچنین، جایگزینی pseudo-labeling در LEPID و SSALD با برچسب‌گذاری مبتنی بر هسته RBF، موجب کاهش خطای انتشار برچسب و استفاده از نظارت احتمالاتی در شرایط رانش مفهوم می‌شود. در مقایسه با روش سه‌چالشی SSALD-EIS که از oversampling ساده، بافر یا بازتولید میکروکلاستر برای داده‌های نامتوازن استفاده می‌کنند، روش پیشنهادی با تولید نمونه‌های مصنوعی از ترکیب وزن‌دار نمونه‌های اقلیت، نویز کنترل‌شده و دنباله آشوبی Logistic Map، تنوع داده‌ها را افزایش و بیش‌برازش را کاهش می‌دهد. همچنین، RBFNN تطبیقی این روش، ساختار هندسی فضای ویژگی را به‌طور مستقیم مدل می‌کند، که رویکردی پیشرفته‌تر نسبت به روش‌های مبتنی بر یادگیری جمعی، KNN با مدل‌های افزایشی است. برخلاف روش AACE-DI که رانش مفهوم را با تأخیر و پس از وقوع تشخیص می‌دهند، روش پیشنهادی رانش مفهوم را به‌صورت پیوسته مدیریت می‌کند. همچنین در تابع خطا، با ترکیب Focal loss، وزن‌دهی پویای کلاس‌ها و بهینه‌سازی تطبیقی Adam، برخلاف توابع کلاسیک (آنتروپی متقاطع)، تمرکز مدل به‌طور هم‌زمان بر نمونه‌های دشوار، کلاس اقلیت و نویزهای ناشی از رانش معطوف می‌شود.

در مجموعه‌داده Covetype، روش پیشنهادی نسبت به SSALD در تمامی معیارهای ارزیابی بهبود نشان می‌دهد. افزایش هم‌زمان در F1-score و G-Mean نشان‌دهنده بهبود تعادل بین نرخ‌های خطای نوع اول و نوع دوم است، به این معنا که مدل توانسته است بدون افزایش حساسیت بیش از حد به کلاس مثبت، عملکرد کلی طبقه‌بندی را پایدار نگه دارد. از منظر تحلیلی، این رفتار بیانگر بهبود در مرز تصمیم‌گیری است؛ به‌گونه‌ای که نمونه‌های مرزی با دقت بیشتری تفکیک شده‌اند. همچنین افزایش Specificity نشان می‌دهد که نرخ false positive کاهش یافته است که معمولاً ناشی از کاهش هم‌پوشانی توزیع ویژگی‌ها بین کلاس‌ها و بهبود قدرت تفکیک ویژگی‌ها در فضای نهفته است. این موضوع بیانگر آن است که مدل پیشنهادی دچار سوگیری به سمت کلاس غالب نشده و توانسته توازن یادگیری را حفظ کند. در مجموعه‌داده Drebin، روش پیشنهادی نسبت به SSALD در تمامی معیارها و به‌ویژه در F1-score و G-Mean عملکرد بهتری دارد. هم‌زمانی افزایش Sensitivity و Specificity نشان می‌دهد که مدل در هر دو بعد تشخیص مثبت‌های واقعی و اجتناب از مثبت‌های کاذب بهبود یافته است. این رفتار معمولاً نشان‌دهنده کاهش trade-off کلاسیک بین recall و precision است و می‌تواند ناشی از یادگیری ویژگی‌های discriminative قوی‌تر یا بهبود مکانیزم وزن‌دهی

نمونه‌ها در جریان یادگیری باشد. از آنجا که Drebin دارای ساختار ویژگی پیچیده و چندبعدی است، بهبود هم‌زمان این دو معیار نشان‌دهنده پایداری مدل در فضای ویژگی‌های با ابعاد بالا و کاهش اثر نویز ساختاری در داده است. در مجموعه‌داده Sine، روش پیشنهادی نسبت به SSALD بهبود قابل توجهی در تمامی معیارهای ارزیابی نشان می‌دهد. این بهبود را می‌توان به توانایی مدل در سازگاری با نوسانات غیرایستا نسبت داد. داده‌های Sine به‌طور ذاتی دارای ساختار دوره‌ای و تغییرپذیر هستند، بنابراین مدل‌هایی که فاقد مکانیسم تطبیق پویا هستند معمولاً دچار افت عملکرد می‌شوند. افزایش پایدار معیارها نشان می‌دهد که روش پیشنهادی توانسته است تغییرات توزیع را به‌صورت تدریجی جذب کرده و پارامترهای خود را با الگوی داده هماهنگ کند که این امر بیانگر توانایی مناسب در مدیریت رانش مفهوم است. در مقایسه با روش‌های AACE-DI، LEPID و SSALD-EIS، روش پیشنهادی در معیارهای G-Mean و Sensitivity عملکرد برتری دارد. این نتیجه نشان‌دهنده افزایش توان مدل در حفظ عملکرد متوازن بین کلاس‌ها در شرایط عدم‌توازن داده است. به‌طور خاص، بهبود G-Mean نشان می‌دهد که مدل در هر دو کلاس مثبت و منفی عملکرد یکنواخت‌تری دارد و دچار سوگیری به سمت کلاس غالب نشده است. همچنین افزایش Sensitivity بیانگر توانایی بالاتر در شناسایی نمونه‌های مثبت در حضور تغییرات توزیع و نویز احتمالی است که معمولاً نشان‌دهنده توانایی بالاتر الگوریتم در برابر تغییرات تدریجی توزیع داده است. در مجموعه‌داده Electricity، روش پیشنهادی نسبت به SSALD در معیارهای F1-score و Specificity عملکرد بهتری دارد، در حالی که در سایر معیارها اختلاف محدودتر است. این رفتار نشان می‌دهد که مدل در کاهش false positive و بهبود کیفیت پیش‌بینی‌های منفی عملکرد مؤثرتری داشته است. با توجه به ماهیت پویا و غیرایستای این مجموعه‌داده، اختلاف عملکرد محدود در برخی معیارها می‌تواند ناشی از تغییرات سریع توزیع داده و وجود الگوهای زمانی پیچیده باشد. با این حال، بهبود F1-score نشان می‌دهد که مدل توانسته است trade-off بین precision و recall را به‌صورت نسبی بهینه‌سازی کند.

شکل (۱) تغییرات دو معیار G-Mean و F1-score را برای مجموعه‌داده Elec نشان می‌دهد. در این شکل، محور افقی بیانگر تعداد نمونه‌های پردازش‌شده و محور عمودی مقدار معیارهای G-Mean و F1-score است. منحنی سبز مربوط به معیار F1-score و منحنی آبی نشان‌دهنده معیار G-Mean می‌باشد. این دو معیار به‌منظور ارزیابی عملکرد مدل در داده‌های نامتوازن مورد استفاده

در شکل (۳)، تغییرات F1-score و G-Mean در مجموعه داده Sine نشان می‌دهد که افت موقت در حوالی ۳۵۰۰ نمونه بیانگر مواجهه مدل با تغییر ساختاری در الگوی داده است. با توجه به ماهیت غیرایستا و نوسانی این داده، چنین افت‌هایی قابل انتظار است. با این حال، افزایش مجدد معیارها نشان‌دهنده توانایی مدل در بازیابی عملکرد پس از تغییر توزیع و انطباق سریع با الگوهای جدید است که این رفتار مناسب مدل در برابر تغییرات دوره‌ای و غیرخطی داده است.

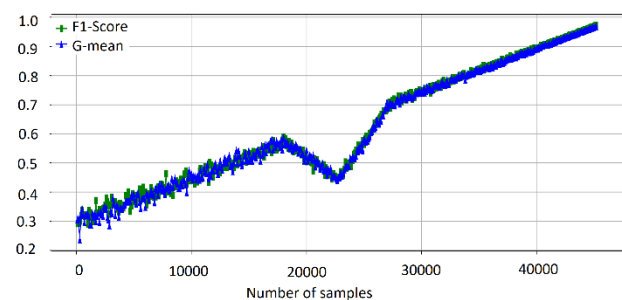


شکل ۳- معیار F1-Score و G-mean روش پیشنهادی با مقادیر مختلف N برای مجموعه داده Sine

۵- نتیجه گیری و کارهای آینده

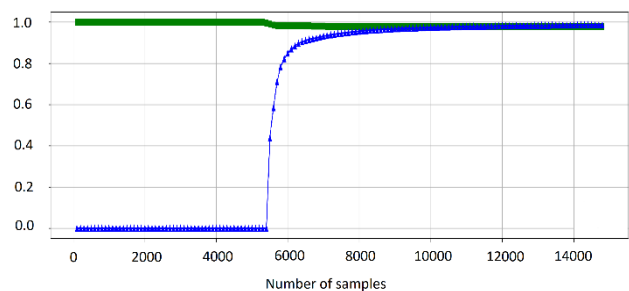
روش پیشنهادی با هدف مقابله با چالش مفهومی در یادگیری نیمه‌نظارتی همراه با رانش مفهوم معرفی شده است. در این روش، مکانیزم یادگیری فعال تطبیقی طراحی شده است که با بهره‌گیری از معیارهای عدم قطعیت و چگالی، نمونه‌های مؤثر و نماینده برای برچسب‌گذاری انتخاب می‌شوند. این مکانیزم موجب افزایش کارایی فرآیند یادگیری و کاهش هزینه‌های برچسب‌گذاری گردیده است. همچنین، در پژوهش حاضر، روش تولید داده‌های مصنوعی وزنی توسعه داده شده است که به بهبود شناسایی کلاس‌های اقلیت کمک می‌کند. در این مکانیزم، با ترکیب هوشمندانه نمونه‌های واقعی و افزودن نویز، تنوع داده‌ها افزایش یافته و تعادل بین کلاس‌ها بهبود یافته است. علاوه بر این، با استفاده از Focal loss، توجه بیشتری به کلاس اقلیت داده شده است. بهبود چشمگیر عملکرد مدل در معیارهای کلیدی مانند F1-Score، G-Mean، Sensitivity و Specificity در مقایسه با روش‌های موجود مشاهده شده است که نشان‌دهنده توانایی بالایی چارچوب پیشنهادی در حفظ تعادل میان شناسایی صحیح نمونه‌های مثبت و منفی در شرایط واقعی است. از منظر کاربردی، چارچوب پیشنهادی می‌تواند در حوزه‌هایی مانند تشخیص نفوذ در شبکه، تشخیص تقلب مالی، تحلیل داده‌های اینترنت اشیا، سیستم‌های پایش پزشکی و سامانه‌های هوشمند صنعتی مورد استفاده قرار

قرار گرفته‌اند. تغییرات معیارهای F1-score و G-Mean در مجموعه داده Electricity نشان می‌دهد که مقادیر اولیه پایین این معیارها ناشی از فاز cold-start و عدم انطباق اولیه مدل با توزیع داده است. با افزایش نمونه‌های پردازش‌شده، روند صعودی هر دو معیار نشان‌دهنده یادگیری تدریجی ساختار داده و کاهش خطای طبقه‌بندی است. افت مشاهده‌شده در بازه ۱۸۰۰۰ تا ۲۰۰۰۰ نمونه می‌تواند ناشی از وقوع رانش مفهوم ناگهانی یا تغییر در توزیع شرطی داده‌ها باشد که موجب افزایش موقت خطای تصمیم‌گیری شده است. با این حال، بازگشت سریع معیارها به مقادیر بالاتر نشان می‌دهد که مدل دارای مکانیزم سازگاری مؤثر بوده و توانسته پارامترهای خود را با توزیع جدید داده تطبیق دهد که بیانگر پایداری دینامیکی الگوریتم است.



شکل ۱- معیار F1-Score و G-mean روش پیشنهادی با مقادیر مختلف N برای مجموعه داده Elec

در شکل (۲)، رفتار دو معیار G-Mean و F1-score را در مجموعه داده Drebin نشان می‌دهد که افزایش سریع این معیارها پس از حدود ۵۰۰۰ نمونه بیانگر عبور مدل از مرحله یادگیری اولیه و دستیابی به نمایش‌های پایدار از داده است. این روند نشان‌دهنده کاهش خطای تعمیم و بهبود تفکیک‌پذیری کلاس‌ها در فضای ویژگی است. همچنین رسیدن معیارها به مقادیر بالا نشان می‌دهد که مدل توانسته است ویژگی‌ها را به‌طور مؤثر استخراج کند و اثر نویز یا هم‌پوشانی کلاس‌ها را کاهش دهد.



شکل ۲- معیار F1-Score و G-mean روش پیشنهادی با مقادیر مختلف N برای مجموعه داده Drebin

- Signal Processing, vol. 38, no. 1, pp. 165–205, 2013, doi: 10.1016/j.ymssp.2013.01.017.
- [9] W. Li, Z. Zhu, F. Jiang, G. Zhou, and G. Chen, “Fault Diagnosis of Rotating Machinery With a Novel Statistical Feature Extraction and Evaluation Method,” *Mechanical Systems and Signal Processing*, vol. 50–51, pp. 414–426, 2015, doi: 10.1016/j.ymssp.2014.05.034.
- [10] F. Lu, Q. Tong, Z. Feng, and Q. Wan, “Unbalanced Bearing Fault Diagnosis Under Various Speeds Based on Spectrum Alignment and Deep Transfer Convolution Neural Network,” *IEEE Transactions on Industrial Informatics*, vol. 19, no. 7, pp. 8188–8198, Jul. 2023, doi: 10.1109/TII.2022.3217541.
- [11] L. Jian, K. Shao, Y. Liu, J. Li, and X. Liang, “OEC: An Online Ensemble Classifier for Mining Data Streams With Noisy Labels,” *Data Mining and Knowledge Discovery*, vol. 38, 2024, doi: 10.1007/s10618-023-00990-0.
- [12] D. Mulimani, P. Patil, S. Totad, and R. Benni, “Online Detection and Adaptation of Concept Drift in Streaming Data Classification,” *Procedia Computer Science*, vol. 235, pp. 2803–2811, 2024, doi: 10.1016/j.procs.2024.04.265.
- [13] Z. Wu, H. Wang, J. Guo, Q. Yang, and J. Shao, “Learning Evolving Prototypes for Imbalanced Data Stream Classification With Limited Labels,” *Information Sciences*, vol. 679, Art. no. 120979, 2024, doi: 10.1016/j.ins.2024.120979.
- [14] J. Li, K. Malialis, C. G. Panayiotou, and M. M. Polycarpou, “Unsupervised Incremental Learning With Dual Concept Drift Detection for Identifying Anomalous Sequences,” in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*, Yokohama, Japan, 2024.
- [15] B. Jiao, H. M. Gomes, B. Xue, Y. Guo, and M. Zhang, “A Semi-Supervised Active Learning Neural Network for Data Streams With Concept Drift,” *IEEE Computational Intelligence Magazine*, vol. 20, no. 1, Feb. 2025, doi: 10.1109/MCI.2024.3486680.
- [16] D. Himaja, V. Dondeti, S. Uppalapati, and S. Virupaksha, “Cluster-Based Active Learning for Classification of Evolving Streams,” *Evolutionary Intelligence*, vol. 17, 2024, doi: 10.1007/s12065-023-00879-3.
- [17] B. Halder, K. M. A. Hasan, T. Amagasa, and M. M. Ahmed, “Autonomic Active Learning Strategy Using Cluster-Based Ensemble Classifier for Concept Drifts in Imbalanced Data Stream,” *Knowledge and Information Systems*, 2025, doi: 10.1007/s10115-025-02647-4.
- [18] S. Li, W. Wang, X. Jiang, J. Hua, J. Mao, Y. Lu, X. Jia, D. Wang, Z. Wang, and Y. Lu, “DDAL: A Drift Detection Active Learning Mechanism,” in *Proc. 11th Int. Conf. Advanced Cloud and Big Data (CBD)*, Ocean Flower Island, Hainan, China, 2023.

پاورقی‌ها:

- ¹ Online Ensemble Classification
- ² Normalized Exponentiated Gradient
- ³ Online Learning from Incomplete and Few Labels
- ⁴ Variational Autoencoder for Anomalous Sequences
- ⁵ Semi-supervised active learning neural network for data streams with concept drift
- ⁶ Cluster based active learning for class imbalance and concept drift
- ⁷ Semi-Supervised Active Learning Ensemble for Imbalanced Streams
- ⁸ Autonomic active learning strategy with a cluster-based
- ⁹ Drift Detection Active Learning
- ¹⁰ Smoothness Assumption
- ¹¹ Adaptive Moment Estimation
- ¹² <https://www.kaggle.com/datasets/uciml/forest-cover-type-dataset>
- ¹³ <https://www.kaggle.com/datasets/sazidthe1/global-electricity-production>
- ¹⁴ <https://drebin.mlsec.org>

گیرد؛ زیرا در این محیط‌ها داده‌ها به‌صورت پیوسته وارد سیستم می‌شوند، توزیع داده‌ها در طول زمان تغییر می‌کند و معمولاً با کمبود داده‌های برچسب‌خورده و داده‌های نامتوازن مواجه هستیم. همچنین، کاهش هزینه برچسب‌گذاری و توانایی سازگاری با رانش مفهوم، امکان استفاده عملی از روش پیشنهادی را در سیستم‌های بلادرنگ و مقیاس‌پذیر فراهم می‌کند. به‌طور کلی، در این پژوهش، با ترکیب راهکارهای یادگیری نیمه‌نظارتی، یادگیری فعال و تطبیقی، مدلی کارآمد، انعطاف‌پذیر و پایدار برای طبقه‌بندی جریان داده‌ها در محیط‌های واقعی ارائه شده است که می‌تواند نقش مهمی در توسعه سیستم‌های هوشمند در حوزه‌های مختلف ایفا کند.

تغییر سریع در توزیع داده‌ها ممکن است موجب ناکارآمد شدن مراکز RBF انتخاب‌شده گردد و فرآیند به‌روزرسانی تطبیقی زمان‌بر شود یا به‌طور کامل نتواند تغییرات را پوشش دهد. به‌عنوان پیشنهاد برای کارهای آینده، چارچوب‌های تشخیص رانش مفهوم مانند ADWIN یا DDM می‌توانند به‌عنوان یک لایه هوشمند نظارتی به الگوریتم پیشنهادی افزوده شوند تا تغییرات احتمالی در توزیع داده‌ها به‌صورت پویا تشخیص داده شوند. این کار باعث می‌شود فرآیند به‌روزرسانی مراکز RBF تنها زمانی انجام شود که واقعاً نیاز باشد.

مراجع

- [1] F. Ceschin, M. Botacin, H. M. Gomes, F. Pinage, L. S. Oliveira, and A. Gregio, “Fast & Furious: On the Modelling of Malware Detection as an Evolving Data Stream,” *Expert Systems with Applications*, vol. 212, Art. no. 118590, 2023, doi: 10.1016/j.eswa.2022.118590.
- [2] S. Wang, L. L. Minku, and X. Yao, “A Systematic Study of Online Class Imbalance Learning With Concept Drift,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 4802–4820, Oct. 2018, doi: 10.1109/TNNLS.2017.2771290.
- [3] Guo, J. Pu, B. Jiao, Y. Peng, D. Wang, and S. Yang, “Online Semi-Supervised Active Learning Ensemble Classification for Evolving Imbalanced Data Streams,” *Applied Soft Computing*, vol. 155, Art. no. 111452, 2024, doi: 10.1016/j.asoc.2024.111452.
- [4] J. Hamidzadeh, N. Kashefi, and M. Moradi, “Instance Reduction Using Multi-Objective Chaotic Particle Swarm Optimization Algorithm,” *Journal of Soft Computing and Information Technology (JSCIT)*, vol. 8, no. 1, pp. 39–49, 2019.
- [5] R. Rastogi and S. Sharma, “Fast Laplacian Twin Support Vector Machine With Active Learning for Pattern Classification,” *Applied Soft Computing*, vol. 74, pp. 343–353, 2019, doi: 10.1016/j.asoc.2018.10.042.
- [6] Y. Hu, P. Baraldi, F. Di Maio, and E. Zio, “A COMPOSE-Based Method for Fault Diagnostics in Evolving Environment,” in *Proc. Prognostics and System Health Management Conf. (PHM)*, 2015.
- [7] Z. Zhang and Q. Yang, “Unsupervised Feature Learning With Reconstruction Sparse Filtering for Intelligent Fault Diagnosis of Rotating Machinery,” *Applied Soft Computing*, vol. 115, Art. no. 108207, 2022, doi: 10.1016/j.asoc.2021.108207.
- [8] Z. Feng, M. Liang, and F. Chu, “Recent Advances in Time-Frequency Analysis Methods for Machinery Fault Diagnosis: A Review With Application Examples,” *Mechanical Systems and*