یادگیری برخط داده‌های جریانی نامتوازن دارای رانش مفهوم به‌وسیلهٔ نظریه باور و تابع آشوب

جواد حمیدزاده* محمدعلی رشیدی محمودی[۲] و منا مرادی[۳]
دانشکده مهندسی کامپیوتر و فناوری اطلاعات، دانشگاه سجاد، مشهد، ایران

چکیده

خصوصیات داده‌های جریانی در گذر زمان ناپایدار بوده و توزیع طبقات متحمل تغییرات می‌شوند؛ بنابراین مدل‌های یادگیری اغلب نیاز به تطبیق با رانش مفاهیم دارند. در این مقاله، با هدف حل دو چالش نبود توازن میان طبقات مشاهده‌شده و وقوع رانش مفهوم، طبقه‌بند داده‌های جریانی نامتوازن دارای رانش مفهوم ارائه شده است. روش پیشنهادی سعی در حذف داده‌های جریانی مرزی و نوفه‌ای با کمک خوشه‌بندی دارد. داده‌ها با کمک تابع باور وزن‌دهی شده است و با درنظرگرفتن برچسب داده‌ها، نمونه‌افزایی در نواحی کم‌تراکم طبقه کمینه و با رویکرد آشوبی انجام می‌گیرد. سپس، با تعریف حدود آستانه، رانش مفهوم پدید آمده از نوع تدریجی و افزایشی شناسایی می‌شود. پیش‌بینی برچسب به‌وسیلهٔ طبقه‌بند ترکیبی و رأی‌گیری وزن‌دار بیشینه انجام می‌پذیرد. عملکرد روش پیشنهادی بر روی مجموعه‌داده‌های پایگاه داده UCI به‌وسیلهٔ روش LOO ارزیابی و با طبقه‌بندهای مرز دانش مقایسه شده‌است. نتایج آزمایش‌ها نشان‌دهندهٔ برتری روش پیشنهادی از نظر معیارهای ارزیابی است.

واژگان کلیدی: تغییر مفهوم، داده جریانی، داده نامتوازن، طبقه‌بندی برخط، نظریه باور.

Online Learning for Imbalanced Data Streams with Concept Drift by Belief Theory and Chaotic Function

Javad Hamidzadeh *, Mohammad Ali Rashidi Mahmoodi and Mona Moradi
Faculty of Computer Engineering and Information Technology,
Sadjad University, Mashhad, Iran

Abstract

Continual learning from data streams is a pivotal aspect of machine learning, requiring the development of algorithms capable of adapting to incoming data. However, the ongoing evolution of data streams presents a formidable challenge as previously acquired knowledge may become outdated. This challenge, known as concept drift, demands timely detection for the effective adaptation of learning models. While various drift detectors have been proposed, they often assume a relatively balanced class distribution. In scenarios with imbalanced data streams, these detectors may exhibit bias toward majority classes, overlooking shifts in minority classes. Moreover, the imbalance among classes can change over time, with roles shifting between majority and minority classes, especially when relationships among classes become complex due to overlapping regions. In this paper, a novel classification method is introduced for imbalanced streaming data affected by concept drift. The proposed method continuously monitors arriving streams to detect and adapt to both imbalances and concept drift. Upon receiving a new block of data, the proposed method employs the k-means clustering approach to identify non-dense regions and performs oversampling for minority classes. Cluster centers are selected using the belief function to address overlapping issues between majority and minority classes. Utilizing a chaotic approach, the new sample is added based on its neighborhood and the size of

* نویسندهٔ عهده‌دار مکاتبات
* Corresponding author

• تاریخ ارسال مقاله: ۱۴۰۰/۴/۱۰ • تاریخ پذیرش: ۱۴۰۲/۴/۱۴ • تاریخ انتشار: ۱۴۰۲/۱۲/۲۹ • نوع مطالعه: پژوهشی

[DOI: 10.61186/jsdp.20.4.23]

thresholds that cover time intervals and classification errors. Finally, the label prediction process is done by ensemble learning and weighted majority voting. Experiments conducted on benchmark datasets from the UCI database evaluate the performance of the proposed method using Leave-One-Out (LOO) validation and comparisons with state-of-the-art methods. The results demonstrate the superiority of the proposed method across various evaluation criteria, highlighting its effectiveness in addressing imbalanced streaming data with concept drift.

Keywords: **Belief Theory; Concept Drift; Data Stream; Imbalanced Data; Online Classification.**

۱- مقدمه

داده‌های جریانی داده‌هایی با حجم بالا هستند که به‌صورت افزایشی و با سرعت بالا به یک سامانه وارد می‌شوند. این داده‌ها با دو چالش مهم مواجه‌اند. چالش نخست، تغییر الگوی توزیع داده‌ها در طول زمان است که به آن رانش مفهوم می‌گویند. طبق رابطه بیز، رانش مفهوم از سه دیدگاه قابل تعریف است: تغییر در احتمال پسین $Pr(y \mid x)$، تغییر در توزیع ویژگی $Pr(x)$ و تغییر در احتمال پیشین طبقه $Pr(y)$. این سه تغییر ممکن است به‌طور هم‌زمان رخ دهند. چالش دوم، عدم توازن در توزیع نمونه‌ها است که به آن داده نامتوازن می‌گویند. داده‌های نامتوازن در بسیاری از مسائل دنیای واقعی نظیر تشخیص نفوذ در شبکه‌های رایانه‌ای، تشخیص بیماری، تشخیص خطا در ماشین‌آلات صنعتی و ... وجود دارند. در مسائل طبقه‌بندی داده‌های نامتوازن، تعداد نمونه‌های طبقه کمینه بسیار کمتر از تعداد نمونه‌های طبقه بیشینه است. بیش‌تر روش‌های سنتی یادگیری ماشین داده‌های جریانی فرض می‌کنند، توزیع همه طبقات متوازن بوده و هزینه یک‌سانی برای وقوع خطا در طبقه‌بندی در نظر می‌گیرند. هرچند این فرض صحیح نبوده و عدم توجه به مسئله توازن در داده‌ها باعث می‌شود که نتایج طبقه‌بندی به نفع طبقه بیشینه تغییر یابند، درحالی‌که هدف بیش‌تر مسائل طبقه‌بندی، شناسایی داده‌های طبقه کمینه است.

در یک دسته‌بندی کلی، روش‌های حل مسئله عدم توازن در طبقات، به دو سطح الگوریتم و سطح داده تقسیم می‌شوند. روش‌های سطح الگوریتم، بر اساس الگوریتم اعمال شده و ویژگی‌های داده نامتوازن، سعی در کاهش حساسیت نمونه‌های طبقه کمینه دارند. روش‌های یادگیری تک‌طبقه، حساس به هزینه و یادگیری جمعی سه روش معروف این دسته هستند. روش‌های یادگیری جمعی برای آموزش داده‌های نامتوازن، چندطبقه‌بند ضعیف را ترکیب می‌کند. این امر منجر به تولید نتایجی با صحت بالا می‌شود. هرچند ایراد این روش آن است که مدت‌زمان آموزش افزایش می‌یابد. این امر مناسب کاربردهای عملی نبوده، همچنین برای داده‌های با ابعاد بالا نیز با محدودیت‌هایی مواجه است.

روش‌های سطح داده با حذف نمونه‌هایی از طبقه بیشینه (نمونه زدایی) و یا تولید نمونه‌های طبقه کمینه (نمونه‌افزایی) سعی در کاهش حساسیت طبقه کمینه دارند. ساده‌ترین رویکردها در سطح داده، نمونه‌افزایی و نمونه‌زدایی تصادفی هستند [۱-۳]. بدین معنی که نمونه‌های هر طبقه با احتمال برابر تولید/حذف می‌شوند؛ هرچند ممکن است، مشکل بیش‌برازش را در پی داشته باشد و یا نمونه‌های نوفه‌ای، مرزی و یا هم‌پوشان با سایر طبقات تولید کند. صحت یادگیری در طبقه نامتوازن، متأثر از دو عامل توزیع طبقه کمینه در داده‌های جریانی و درجه هم‌پوشانی میان طبقات است. یک روش مناسب برای حل این مسئله، خوشه‌بندی نمونه‌های مجموعه‌داده نامتوازن و سپس نمونه‌افزایی است که در [۴, ۵] ارائه شده است. هر چند ایراد این روش‌ها درنظرنگرفتن طبقه واقعی نمونه‌های مجموعه‌داده است.

مسئله طبقه‌بندی داده‌های جریانی، در حضور رانش مفهوم و داده نامتوازن مسئله جدید و پرکاربردی بوده و تعداد اندکی از مطالعات تاکنون به حل هم‌زمان دو چالش پرداخته‌اند. بر اساس تعداد نمونه‌های پردازش‌شده در هر لحظه، روش‌های موجود به دو دستهٔ: (۱) برخط و (۲) مبتنی بر قطعه تقسیم می‌شوند. در روش‌های برخط، در هر لحظه تنها یک داده پردازش شده، بنابراین در صورت ایستایی محیط، کارایی پایینی دارند. روش‌های مبتنی بر قطعه در هر لحظه، از دسته‌ای از داده‌ها برای آموزش مدل یادگیر استفاده کرده است؛ بنابراین کارایی به‌نسبه باثباتی داشته اما توانایی مدیریت رانش مفهوم از نوع تغییر در احتمال پیشین‌طبقه $Pr(y)$ را ندارند [۶]. این مسئله در صورت نبود توازن در داده‌های جریانی و به‌خصوص در شرایطی که نمونه‌های طبقه کمینه به طبقه بیشینه تعلق می‌یابند، مشکل جدی ایجاد می‌کند.

مقاله حاضر با درنظرگرفتن این دو مسئله، روشی نوین برای بهبود فرآیند یادگیری این نوع از داده‌ها ارائه کرده است. روش پیشنهادی در گام نخست از روش ابداعی نمونه‌افزایی آشوبی با استفاده از همسایه متقابل (CMNOS[1]) برای حل مشکل عدم توازن استفاده می‌کند.

[1] Chaotic Mutual Neighbor-based Over-Sampling

[DOI: 10.61186/jsdp.20.4.23]

در این روش، پس از خوشه‌بندی مجموعه‌داده، با بهره‌گیری از تابع باور، به هر نمونه وزن مناسب تخصیص داده شده و سپس با استفاده از تابع آشوب و در نظر گرفتن میزان تراکم حول هر نمونه از طبقه کمینه، نمونه‌افزایی صورت می‌گیرد. در گام دوم، با محاسبه خطای طبقه‌بند و دو حد آستانه خطا و یک حد آستانه زمانی، وقوع رانش مفهوم شناسایی و با استفاده از رویکرد یادگیری جمعی، برچسب هر داده تعیین می‌شود.

دستاوردهای روش پیشنهادی عبارت‌اند از:

- حل هم‌زمان دو مسئله رانش مفهوم و عدم توازن طبقات در داده‌های جریانی
- وزن‌دهی به نمونه‌ها توسط تابع باور
- نمونه‌افزایی هدفمند با استفاده از تابع آشوب در نواحی کم تراکم
- تشخیص رانش مفهوم ناگهانی و تدریجی با استفاده از دو حد آستانه خطا و یک حد آستانه زمانی

ساختار مقاله بدین شرح است: در بخش دوم، کارهای پیشین بررسی و در بخش سوم مفاهیم اولیه رانش مفهوم و نظریه باور شرح داده شده است. در بخش چهارم روش پیشنهادی و در بخش پنجم آزمایش‌ها و تفسیر آن‌ها آورده شده است. بخش ششم به نتیجه‌گیری و کارهای آینده اختصاص دارد.

۲-کارهای پیشین

تاکنون روش‌های مختلفی برای غلبه بر رانش مفهوم ارائه شده است. در یک دسته‌بندی کلی این روش‌ها به دو دسته تقسیم می‌شوند: (۱) روش‌هایی که در آن مدل یادگیر بدون توجه به وقوع تغییر مفهوم، به طور مرتب در بازه‌های مشخصی اقدام به به‌روزرسانی می‌کند. (۲) روش‌هایی که در آن مدل یادگیر تنها در صورت وقوع تغییر مفهوم اقدام به به‌روزرسانی می‌کند. این روش‌ها با پایش شاخص‌هایی نظیر کیفیت عملکرد مدل یادگیر و یا خصوصیات داده از وقوع تغییر مطلع می‌شوند.

در یک دسته‌بندی دیگر، روش‌های شناسایی رانش مفهوم به چهار دسته تقسیم می‌شوند. دسته نخست متعلق به روش‌های مبتنی بر تحلیل توالی داده است. این روش‌ها به طور پیوسته نتایج پیش‌بینی را بررسی می‌کنند [۷, ۸]. دسته دوم متعلق به روش‌های مبتنی بر تحلیل پارامترهای آماری بر اساس روش‌های آماری نظیر مقدار میانگین و انحراف معیار نتایج پیش‌بینی است [۹]. روش‌های DDM [۱۰] و EDDM [۱۱] با مقایسه خطای طبقه‌بند رانش مفهوم را شناسایی می‌کنند. روش‌های دسته سوم مبتنی بر پنجره هستند. در این روش‌ها، از دو پنجره ثابت و لغزان استفاده می‌شود. پنجره ثابت حاوی اطلاعات خلاصه‌شده‌ای از داده‌های قدیمی و پنجره لغزان دربرگیرنده داده‌های جدید است. در صورت مشاهده تفاوت چشم‌گیر میان توزیع این دو پنجره، می‌توان وقوع رانش مفهوم را نتیجه‌گیری کرد. برخلاف اینکه روش‌های دسته سوم نتایج دقیق‌تری را تولید می‌کند، اما ایراد آن‌ها این است که به‌دلیل مقایسه توزیع داده‌ها با روش‌های آماری، نیازمند زمان و حافظه بیشتری هستند. روش‌هایی نظیر ADWIN [۱۲]، الگوریتم VFDT [۱۳] و انواع بهبودیافته آن [۱۴, ۱۵] درخت تصمیم افزایشی برای یادگیری داده‌های جریانی ارائه می‌دهد. روش‌هایی مانند HDDM [۱۶]، FHDDM [۱۷] و FHDDMS [۱۸] از روش‌های شناخته‌شده این دسته هستند. در [۱۹] روش بدون ناظر و با اعمال معیار آنتروپی اطلاعات در خوشه‌بندی k-means جهت شناسایی رانش مفهوم داده‌های جریانی ارائه شده‌است. روش‌های مبتنی بر یادگیری جمعی در دسته چهارم قرار می‌گیرند [۶, ۲۰, ۲۱]. ازآنجاکه این روش‌ها از چند روش یادگیری به‌طور هم‌زمان برای تصمیم‌گیری استفاده می‌کنند، اغلب توانایی بالایی در حل مسائل پیچیده دارند. استفاده از روش‌های مبتنی بر یادگیری جمعی این مزیت را دارد که به هنگام مواجهه با تعداد داده بسیار زیاد می‌توان با کمک ترکیب چند مدل با خطای طبقه‌بندی متفاوت، کل فضای ویژگی را شناسایی کرد. مسئله مهم در این روش‌ها، انتخاب اندازه مناسب پنجره است. پنجره با اندازه کوچک به‌سرعت تغییرات را شناسایی می‌کند و این تضمین را می‌دهد که مدل یادگیر به‌سرعت با تغییرات منطبق شود؛ اما ممکن است در طولانی‌مدت و با استمرار وقوع تغییرات در داده‌های جریانی، عملکرد مدل را تحت تأثیر قرار دهد. از سویی دیگر، پنجره با اندازه بزرگ، امکان یادگیری پایدار را فراهم نموده اما مدل قادر به بروز واکنش مناسب و سریع در مواجهه با تغییرات ناگهانی نیست. به هنگام پایش و درصورت وقوع رانش مفهوم، مدل یادگیر باید با تنظیم اندازه پنجره با تغییرات انطباق یابد. به‌صورت یک قاعده کلی، اندازه پنجره با وقوع رانش مفهوم کاهش و در غیر این صورت افزایش می‌یابد.

تاکنون مطالعات گوناگونی درخصوص حل هم‌زمان عدم توازن طبقات و وقوع رانش مفهوم انجام شده، هرچند هنوز نیازمند بررسی بیشتر است [۲۲-۲۶]. در [۲۲] روش یادگیری فعال برای طبقه‌بندی داده‌های جریانی با چند طبقه نامتوازن در حضور رانش مفهوم ارائه شد.

[DOI: 10.61186/jsdp.20.4.23]

روش برخط یادشده از یادگیری جمعی برای تصمیم‌گیری در مورد برچسب داده‌های ورودی استفاده می‌کند. در [۲۳] رویکرد یادگیری جمعی همراه با روش‌های نمونه‌افزایی و نمونه زدایی برای تخمین برچسب داده‌های جریانی ارائه گردید. طبقه‌بندهای موجود در یادگیری جمعی به صورت پویا انتخاب می‌شوند. در [۲۷] رویکرد یادگیری جمعی ارائه شد که پیش از یادگیری هر پنجره ورودی، نمونه زدایی با استفاده از الگوریتم *k*-means انجام می‌شود. روش [۲۸] با رویکردی مشابه سه الگوریتم SERA، MUSERA و REA را ارائه نمود که نمونه‌های منتخب طبقه کمینه را برای نمونه‌افزایی به پنجره جاری اضافه می‌کنند. در [۲۹] روشی برای محاسبه وزن طبقه‌بندهای آموزش‌دیده بر روی پنجره جاری ارائه شد. در [۳۰] با تغییر Learn++ دو الگوریتم Learn++.NIE و Learn++.CDS برای حل مشکل عدم توازن داده‌های جریانی ارائه شد که برخلاف دست‌یابی به نتایج مطلوب، دو الگوریتم دارای بار محاسباتی بالایی هستند. در [۶] روش مبتنی بر یادگیری افزایشی برای حل هم‌زمان دو مسئله عدم توازن و رانش مفهوم ارائه شد. این روش برای حل مسئله عدم توازن از نمونه‌افزایی تصادفی استفاده می‌کند.

۳- مفاهیم اولیه

این بخش به تشریح مبانی دو مفهوم به‌کاررفته در روش پیشنهادی می‌پردازد. در بخش ۳-۱ رانش مفهوم و در بخش ۳-۲ تابع باور بیان شده‌اند.

۳-۱- رانش مفهوم

فرض کنید داده جریانی توالی از (x_i, y_i) برای $i = 1, 2, \ldots$ بوده که x برداری m بعدی و y برچسب داده متعلق به مجموعه‌ای شامل تعداد L برچسب است. درصورت وقوع رانش مفهوم توزیع داده‌ها در گذر زمان تغییر می‌کند. به بیانی دیگر، رانش مفهوم بین دو برچسب زمانی متوالی t و $t+1$ به صورت زیر تعریف می‌شود:

$$\exists x : Pr_t(x, y) \neq Pr_{t+1}(x, y) \qquad (۱)$$

که Pr_t توزیع احتمال توأم متغیرهای ورودی x و برچسب y در زمان t است.

در نظریه تصمیم بیز، تصمیم‌گیری در مورد طبقه نمونه به‌صورت زیر انجام می‌شود:

$$Pr(y \mid x) = \frac{Pr(x \mid y) Pr(y)}{Pr(x)} \qquad (۲)$$

که $Pr(y)$ احتمال پیشین طبقه y، $Pr(x \mid y)$ تابع چگالی احتمال مشروط به طبقه و $Pr(x) = \sum_{i=1}^{L} Pr(y_i) Pr(x \mid y_i)$ است. با درنظرگرفتن رابطه **Error! Reference source not found.** می‌توان انواع رانش مفهوم را به دو دسته کلی تقسیم کرد:

- رانش مفهوم واقعی: در این حالت، با تغییر در $Pr(x \mid y)$، مرز تصمیم‌گیری $Pr(y \mid x)$ نیز تغییر می‌کند .
- رانش مفهوم مجازی: در این حالت، توزیع احتمال ویژگی‌های متغیر ورودی x، $Pr(x)$، تغییر نموده اما منجر به تغییر مرز تصمیم‌گیری $Pr(y \mid x)$ نمی‌شود.

همچنین از نظر سرعت وقوع تغییرات، انواع رانش مفهوم به پنج دسته تقسیم می‌شود:

- رانش مفهوم تدریجی: در این نوع رانش، یک تغییر هموار و تدریجی از مفهومی به مفهوم دیگر رخ می‌دهد. مفهوم میانی تولیدشده یکی از مفاهیم آغازین و یا پایانی است.
- رانش مفهوم ناگهانی: در این نوع رانش، یک تغییر ناگهانی در محتوای طبقه رخ داده و طبقه دیگری ظاهر می‌شود.
- رانش مفهوم افزایشی: در این نوع رانش، ضمن تغییر از مفهومی به مفهوم دیگر، چندین مفهوم میانی دیگر ظاهر می‌شوند. مفاهیم میانی تولیدشده ترکیبی از مفاهیم آغازین و پایانی هستند.
- رانش مفهوم بازگشتی: در این نوع رانش، پس از گذشت زمان، مفاهیم از قبل مشاهده‌شده مجدد ظاهر خواهند شد.
- رانش کوتاه: در این نوع رانش، داده پرت با مفهوم موجود آمیخته می‌شود. در این حالت، تغییر مفهوم موقتی بوده و داده‌های بعدی را تغییر نخواهد داد.

خطای پیش‌بینی توالی[1]: فرض کنید توالی از داده‌های جریانی (x_i, y_i) برای $i = 1, 2, \ldots$ برای آزمایش طبقه‌بند (قبل از آموزش آن) در اختیار است. خطای طبقه‌بندی مجموع مقادیر تابع هزینه $f_{Loss}(.)$ میان برچسب پیش‌بینی‌شده $\hat{y}_i$ و برچسب واقعی y_i است که از رابطه زیر به دست می‌آید:

$$p_i = \frac{1}{i} \sum_{i=1}^{j} f_{Loss}(y_i, \hat{y}_i) \qquad (۳)$$

نکته مهم این است که در شرایط ایستایی مسئله و پایایی توزیع داده‌های ورودی، با افزایش تعداد نمونه‌های

[1] Prequential Error

[DOI: 10.61186/jsdp.20.4.23]

آموزشی، خطای پیش‌بینی توالی کاهش یافته و به مقدار بهینه همگرا می‌شود. از طرفی دیگر، افزایش نرخ خطا به معنی پویایی توزیع داده ورودی بوده که می‌تواند به عنوان نشانه‌ای از وقوع رانش مفهوم تلقی شود.

۳-۲- نظریه باور

نظریه توابع باور یا دمپستر-شفر به عنوان تعمیمی از نظریه احتمال، روشی مناسب برای حل مسائلی است که با عدم قطعیت مواجه‌اند [۳۱]. فرض می‌کنیم طبقات مسئله توسط مجموعه متناهی چارچوب تشخیص[1] $\Theta=\{\theta_1,...,\theta_L\}$ تعریف می‌شوند. تابع $m:2^\Theta\to[0,1]$ را تابع جرم گویند اگر دارای شرایط زیر باشد:

$$m(\varnothing)=0 \qquad (۴)$$
$$\sum_{A\subseteq 2^\Theta} m(A)=1$$

که در آن 2^Θ تمام زیرمجموعه‌های ممکن Θ است. به‌عنوان‌مثال، با فرض دودویی‌بودن مسئله، $\Theta=\{\theta_1,\theta_2\}$ بوده و $2^\Theta=\{\varnothing,\theta_1,\theta_2,\theta_1\cup\theta_2\}$ است. $m(A)$ جرم A بوده که درجه اطمینان از رخداد پیشامد A (و نه هیچ زیرمجموعه آن) است. هر مجموعه $A\in 2^\Theta$ که برای آن $m(A)>0$ باشد، یک عضو کانونی m نامیده می‌شود. اعضای کانونی زیرمجموعه‌ای از چارچوب تشخیص هستند که اطلاعات و شواهد موجود، بر آن‌ها متمرکز است. در نظریه توابع باور، دو تابع مهم باور و خشنودی[2] از تابع جرم m به دست می‌آیند. تابع باور بر روی Θ به صورت $Bl:2^\Theta\to[0,1]$ تعریف می‌شود؛ اگر $Bl(A)=\sum_{B\subseteq A}m(B)=1$ که $A\subseteq 2^\Theta$ است. تابع خشنودی بر روی Θ به‌صورت $Pl:2^\Theta\to[0,1]$ تعریف می‌شود؛ اگر $Pl(A)=1-Bl(\bar{A})$. توزیع احتمال A سازگار با تابع جرم m است اگر به ازای $\forall A\subseteq 2^\Theta$ در شرط $Bl(A)<Pr(A)<Pl(A)$ صدق کند. همچنین، توزیع احتمال شرط‌بندی $BetP$ اندازه احتمال برای اتخاذ تصمیم بر روی θ_i را به‌صورت زیر محاسبه می‌کند:

$$BetP(\theta_i)=\sum_{A\subseteq 2^\Theta|\theta_i\in A}\frac{m(A)}{|A|} \qquad (۵)$$

۴- روش پیشنهادی

این بخش به معرفی طبقه‌بند پیشنهادی با رویکرد یادگیری جمعی و افزایشی برای داده‌های جریانی نامتوازن و دارای رانش مفهوم می‌پردازد. هر نمونه داده جریانی در لحظه t به‌صورت $\boldsymbol{x}_t=\{x_1,x_2,...,x_d\}$ که $\{\boldsymbol{x}_t,y_t\}$ بردار ویژگی نمونه d-بعدی و $y\in\{+1,-1\}$ برچسب طبقه است. نمونه‌های با برچسب $y=-1$ متعلق به طبقه بیشینه و نمونه‌های با برچسب $y=+1$ متعلق به طبقه کمینه هستند. در روش پیشنهادی فرض می‌شود که تغییر مفهوم می‌تواند به دو صورت تدریجی و یا به‌صورت ناگهانی رخ دهد. رویکرد روش پیشنهادی به‌صورت زیر است:

در گام نخست، جریان داده ورودی به قطعاتی با اندازه ثابت تقسیم می‌شوند. B_t قطعه داده در زمان جاری t بوده که حاوی دو مجموعه‌داده آموزشی Tr_t و آزمایشی Te_t است. مجموعه‌داده آموزشی $Tr_t=\{\boldsymbol{x}_{Tr_i},y_{Tr_i}\}_{i=1}^{a}$ حاوی a نمونه و مجموعه‌داده آزمایشی $Te_t=\{\boldsymbol{x}_{Te_i}\}_{i=1}^{b}$ حاوی b نمونه بوده به‌گونه‌ای که $|B_t|=a+b$ است. مجموعه‌داده Tr_t در ابتدا نامتوازن بوده و شامل دو طبقه کمینه و بیشینه است. داده‌های آموزشی l امین قطعه داده با روش پیشنهادی به نام CMNOS متوازن شده و برای آموزش طبقه‌بندها مورداستفاده قرار می‌گیرند. رویکرد یادگیری جمعی، از k طبقه‌بند آموزش‌دیده $h_{l,j}$ بر روی l امین قطعه داده $(1\le l<t,1\le j\le k)$ استفاده می‌کند. در گام دوم، وزن هر یک از طبقه‌بندهای حاضر در یادگیری جمعی با درنظرگرفتن تعداد نمونه‌های غلط طبقه‌بندی شده‌شان محاسبه شده و درصورت به حدنصاب رسیدن تعداد طبقه‌بندهای شرکت‌کننده در یادگیری جمعی، طبقه‌بند با بیشترین خطای طبقه‌بندی حذف و طبقه‌بند مبتنی بر یادگیری جمعی به‌روزرسانی می‌شود. در پایان، برچسب مجموعه‌داده‌های آزمایشی Te_t تخمین زده می‌شوند. گفتنی است هر کدام از طبقه‌بندهای حاضر در یادگیری جمعی با زیرمجموعه‌ای از Tr_t آموزش می‌بینند. در ادامه، ابتدا راه‌حل پیشنهادی برای مسئله عدم توازن در توزیع داده‌ها و سپس چگونگی تشخیص رانش مفهوم و پیش‌بینی برچسب داده‌های جریانی بیان شده‌اند.

۴-۱- متوازن‌سازی توزیع طبقات

در داده‌های جریانی نامتوازن، فرایند یادگیری تحت تأثیر توزیع طبقه بیشینه و درجه هم‌پوشانی میان طبقات قرار می‌گیرد. جهت کاهش تأثیر این موارد، در روش پیشنهادی از راه‌کار خوشه‌بندی و نمونه‌افزایی طبقه کمینه مسئله استفاده شده است. ایده اصلی روش ابداعی CMNOS (نمونه‌افزایی آشوبی با استفاده از همسایه متقابل) برای مقابله با عدم توازن و ایجاد توزیع یک‌نواخت

[1] Frame of Discernment
[2] Plausibility Function

[DOI: 10.61186/jsdp.20.4.23]

میان طبقات آن است. در این روش، نمونه‌افزایی برای داده‌هایی که در نواحی کم‌تراکم حضور دارند انجام می‌گیرد. بدین منظور، ابتدا با کمک رویکرد بدون ناظر و با استفاده از الگوریتم خوشه‌بندی k-means داده‌های آموزشی طبقه کمینه خوشه‌بندی و k همسایه هر زوج دوتایی از نمونه‌ها $(\boldsymbol{x}_i, \boldsymbol{x}_j)$ شناسایی می‌شوند. در روش پیشنهادی، مقدار k برابر با صد در نظر گرفته شده‌است. برای محاسبه فاصله از معیار اقلیدسی (رابطه **Error! Reference source not found.**) استفاده شده است:

$$Dist(\boldsymbol{x}_i, \boldsymbol{x}_j) = \sqrt{\sum_{D=1}^{d} (x_i^D - x_j^D)^2} \quad (۶)$$

که منظور از $\boldsymbol{x}_i^D$، بعد D ام نمونه i ام است. پس از شناسایی k همسایه، اگر:

- $\boldsymbol{x}_i$ در k همسایگی $\boldsymbol{x}_j$ بوده ولی $\boldsymbol{x}_j$ در k همسایگی $\boldsymbol{x}_i$ نباشد، آنگاه نمونه‌افزایی حول نمونه $\boldsymbol{x}_j$ و طبق رابطه ***Error! Reference source not found.*** انجام می‌شود.
- $\boldsymbol{x}_j$ در k همسایگی $\boldsymbol{x}_i$ بوده ولی $\boldsymbol{x}_i$ در k همسایگی $\boldsymbol{x}_j$ نباشد، آنگاه نمونه‌افزایی حول نمونه $\boldsymbol{x}_i$ و طبق رابطه **Error! Reference source not found.** انجام می‌شود.
- در سایر حالات، مکان نمونه‌افزایی طبق رابطه **Error! Reference source not found.** تعیین می‌شود.

$$\xi_{r+1} = a\xi_r(1-\xi_r) \quad (۷)$$

$$\boldsymbol{x}' = \boldsymbol{x}_j + \xi Dist(i, j) \quad (۸)$$

$$\boldsymbol{x}' = \boldsymbol{x}_i + \xi Dist(i, j) \quad (۹)$$

$$\boldsymbol{x}' = \boldsymbol{x}_i \pm \xi Dist(i, j) \quad (۱۰)$$

در رابطه **Error! Reference source not found.**، r شماره تکرار و نرخ رشد a برابر با ۴ در نظر گرفته شده است. در روابط بالا، علامت پریم به معنای مکان نمونه جدید در فضای ویژگی d-بعدی است.

ممکن است نمونه‌های جدید در ناحیه طبقه بیشینه ایجاد شوند و یا حتی در نواحی پرت قرار گرفته و به‌عنوان نوفه تلقی شوند. برای تعیین طبقه نمونه جدید و به منظور اطمینان از این مسئله که نمونه‌های ایجادشده در ناحیه طبقه کمینه قرار خواهند گرفت (به فضای ویژگی مشابهی تعلق داشته) و دور از طبقه بیشینه هستند، لازم است فاصله نمونه جدید $\boldsymbol{x}'$ از مراکز طبقه $C = \{\boldsymbol{c}_i\}_{i=1}^{L}$ محاسبه شود. با فرض در اختیار داشتن L طبقه که توسط مجموعه چارچوب تشخیص در تابع باور $\Theta = \{\theta_1, ..., \theta_L\}$، مرکز طبقه $\boldsymbol{c}$ به‌صورت زیر محاسبه می‌شود:

$$\boldsymbol{c}_j = \frac{\sum_{\boldsymbol{x}_i \in \theta_j} \boldsymbol{x}_i}{Z_j} \quad (۱۱)$$

که در آن Z_j تعداد نمونه‌های آموزشی متعلق به طبقه θ_j است. ممکن است نمونه جدید در نواحی هم‌پوشان میان دو یا چند طبقه ایجاد و موجب ابهام شود. ازآنجاکه این ناحیه (θ_U)، اجتماعی از چند طبقه است، یعنی $\theta_U = \theta_i ... \cup \theta_j$، پس مرکز ناحیه هم‌پوشان طبقات درگیر ($\boldsymbol{c}_U$) فاصله‌ای یکسان با هر یک از طبقات دارد؛ به‌عبارتی‌دیگر، $Dist(\boldsymbol{c}_U, \boldsymbol{c}_i) = ... = Dist(\boldsymbol{c}_U, \boldsymbol{c}_j)$ است که $Dist(\boldsymbol{c}_U, \boldsymbol{c}_i)$ طبق فاصله نرمال شده اقلیدسی و از رابطه **Error! Reference source not found.** به دست می‌آید:

$$Dist(\boldsymbol{c}_U, \boldsymbol{c}_i) = \sqrt{\sum_{D=1}^{d} \frac{(\boldsymbol{c}_U^D - \boldsymbol{c}_i^D)^2}{(\delta_i^D)^2}} \quad (۱۲)$$

که δ_i^D انحراف معیار نمونه‌های آموزشی متعلق به طبقه θ_i در بعد D ام است.

برای تشخیص طبقه (برچسب) نمونه جدید $\boldsymbol{x}'$، جرم آن طبق رابطه **Error! Reference source not found.** محاسبه می‌شود:

$$m(\theta_i) = e^{-Dist(\boldsymbol{x}', \boldsymbol{c}_i)} \quad (۱۳)$$

با مقایسه مقادیر به‌دست‌آمده، $\boldsymbol{x}'$ به طبقه‌ای منتسب خواهد شد که جرم بیشتری دارد. همچنین، از آنجا که CMNOS رویکرد نمونه‌افزایی دارد؛ بنابراین، تنها در صورتی نمونه $\boldsymbol{x}'$ به مجموعه‌داده آموزشی افزوده می‌شود که برچسب آن متعلق به طبقه کمینه باشد.

۴-۲- شناسایی رانش مفهوم و پیش‌بینی برچسب داده جریانی

وقوع رانش مفهوم با بهره‌گیری از خطای طبقه‌بندی تشخیص داده می‌شود. پس از ورود داده آزمایشی فاقد برچسب در قطعه داده $Te_t = \{\boldsymbol{x}_{Te_i}\}_{i=1}^{b}$، طبقه‌بندهای پایه طبقه‌بند ترکیبی اقدام به تخمین برچسب آن می‌کنند. در این مقاله، SVM به‌عنوان طبقه‌بند پایه در نظر گرفته شده است. در روش پیشنهادی با هدف تأکید بر نمونه‌های غلط طبقه‌بندی‌شده، وزن $D_t(i)$ بیشتری به آن‌ها داده می‌شود. با فرض اینکه اندکی پس از پیش‌بینی برچسب، برچسب واقعی داده آزمایشی آشکار می‌شود می‌توان خطای قطعه داده را تخمین زد. پیش از آموزش طبقه‌بندهای پایه جدید، توزیع خطای طبقه‌بند ترکیبی

[DOI: 10.61186/jsdp.20.4.23]

فعلی محاسبه می‌شود. ازآنجاکه توزیع خطای قطعه ورودی نامعلوم است، فرض می‌شود از توزیع یکنواخت پیروی نموده پس $D_t(i)=1/b$ است. اگر $h_n(\boldsymbol{x}_{Te_i})$ نتیجه پیش‌بینی برچسب نمونه آزمایشی $\boldsymbol{x}_{Te_i}$ به‌وسیلهٔ طبقه‌بند پایه n ام باشد، خطای طبقه‌بند یادشده از رابطه **Error! Reference source not found.** محاسبه می‌شود:

$$\epsilon_n^t(\boldsymbol{x}_{Te_i})=\sum_{i=1}^{b} D_t(i)\,|\,h_n(\boldsymbol{x}_{Te_i})\neq y_i\,| \qquad (۱۴)$$

درصورتی‌که مقدار حاصل مساوی یا بیشتر از حد آستانه از قبل تعریف‌شده τ_2 باشد، توزیع خطا D_t با استفاده از رابطه **Error! Reference source not found.** به‌روزرسانی شده، همچنین طبقه‌بند مجدداً آموزش می‌بیند.

$$D_t(i)=\begin{cases}\dfrac{1-\epsilon_n^t(\boldsymbol{x}_{Te_i})}{\epsilon_n^t(\boldsymbol{x}_{Te_i})} & \text{if } h_n(\boldsymbol{x}_{Te_i})\neq y_i\\ 1 & \text{Otherwise}\end{cases} \qquad (۱۵)$$

وزن طبقه‌بند پایه n ام w_n^t با استفاده از تابع سیگموئید و به صورت زیر محاسبه می‌شود:

$$\mathrm{sig}_n^t=\frac{1}{1+e^{-r(t-n-s)}},\qquad r,s\in\mathrm{R} \qquad (۱۶)$$

$$w_n^t=\begin{cases}1 & t=n\\ \dfrac{sig_n^t}{sig_n^t+\sum_{j=1}^{t-1} w_n^{t-j}} & \text{Otherwise}\end{cases} \qquad (۱۷)$$

لازم به ذکر است که طبقه‌بند پایه n ام بر روی قطعه داده t ام، با وزن اولیه $w_n^t=1$ شروع به کار می‌کند . در رابطه **Error! Reference source not found.**، مقادیر r و s ثابت بوده و برابر با یک تنظیم شده‌اند.

میزان تأثیر طبقه‌بند پایه n ام در رأی‌دهی $V\_w_n^t$ از رابطه زیر محاسبه می‌شود:

$$V\_w_n^t=\mathrm{Ln}\left(\frac{1}{\sum_{j=1}^{t} w_n^t\epsilon_n^t}\right) \qquad (۱۸)$$

در پایان، برچسب نمونه آزمایشی $\boldsymbol{x}_{Te_i}$ توسط طبقه‌بند ترکیبی H تخمین زده می‌شود. بدین ترتیب، با استفاده از رابطه **Error! Reference source not found.**، $\hat{y}_i$ به دست می‌آید:

$$\hat{y}_i=\arg\max_{\theta}\sum_{n} w_n^t.|h_n(\boldsymbol{x}_i)=\theta| \qquad (۱۹)$$

در روش پیشنهادی، با استفاده از خطای محاسبه‌شده و با درنظرگرفتن سه وضعیت مختلف، وقوع یا عدم وقوع رانش مفهوم نتیجه‌گیری می‌شود:

- وضعیت ۱: اگر $\epsilon_n^t(\boldsymbol{x}_{Te_i})>\tau_1$ باشد، تنها یک اخطار ثبت شده و بدون اعمال به‌روزرسانی طبقه‌بندها (آموزش مجدد و تعیین وزن جدید)، سیستم در انتظار ورود قطعه داده بعدی می‌ماند. τ_1 حد آستانه اول جهت اخطار بوده و با مقدار ۰.۲ تنظیم شده است.
- وضعیت ۲: اگر $\epsilon_n^t(\boldsymbol{x}_{Te_i})>\tau_2$ باشد، وقوع رانش مفهوم تشخیص داده شده، بنابراین به‌روزرسانی طبقه‌بندها انجام می‌شود. τ_2 حد آستانه دوم بوده و با مقدار ۰.۳ تنظیم شده است.
- وضعیت ۳: اگر $\tau_1<\epsilon_n^t(\boldsymbol{x}_{Te_i})<\tau_2$ باشد، برای تشخیص وقوع رانش مفهوم از حد آستانه زمانی τ_T استفاده می‌شود. اگر در τ_T از قبل تعریف‌شده، خطای طبقه‌بندی پایدار بوده و یا افزایش یابد، وقوع رانش مفهوم تشخیص داده می‌شود. τ_T با مقدار ۳ تنظیم شده است. برای درک بهتر، الگوریتم و روندنمای روش پیشنهادی در ادامه نشان داده شده‌اند.

(الگوریتم-۱) تشخیص رانش مفهوم و پیش‌بینی برچسب
(Algorithm-1) Concept drift detection and label prediction

Input: *Block* B_t, *which contains training set* $Tr_t=\{\boldsymbol{x}_{Tr_i},y_{Tr_i}\}_{i=1}^{a}$ *and test set* $Te_t=\{\boldsymbol{x}_{Te_i}\}_{i=1}^{b}$.

Output: *Predicted label* $\hat{y}_i$ *for test sample* $\boldsymbol{x}_{Te_i}$.

Do for $B_t, t=1,2,\ldots$

1. Initialize $D_t(i)=1/b$, $n=t$ *(n: The number of classifiers)*

2. Call base classifier h_n

3. Evaluate the base classifiers by using Eq. (14)

if $\epsilon_n^t(\boldsymbol{x}_{Te_i})>\tau_1$

alarm_flag=1

else if $\epsilon_n^t(\boldsymbol{x}_{Te_i})>\tau_2$ *and alarm_flag=1*

Drift_flag=1

else if $\epsilon_n^t(\boldsymbol{x}_{Te_i})>\tau_1$ *and time* $>\tau_T$

Drift_flag=1

if Drift_flag=1

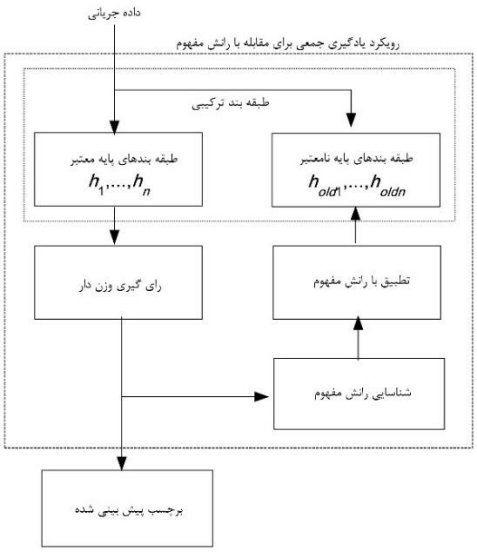
generate new base classifier

4. Calculate the weight of each classifier by using Eq. (17)

5. Calculate the voting weight of each classifier by using Eq. (18)

6. Calculate the predicted label by using Eq. (19)

[DOI: 10.61186/jsdp.20.4.23]



(شکل -۱) روندنمای روش پیشنهادی

(Figure-1) The flowchart of the proposed method

۵- آزمایش‌ها

در این بخش، نتایج حاصل از پیاده‌سازی روش پیشنهادی بر روی سیستم کامپیوتری با پردازنده اینتل ۲.۷۰ گیگاهرتز و ۸ گیگابایت رم و با نرم‌افزار MATLAB R2016 بر روی AA مجموعه‌داده مستخرج از پایگاه داده UCI بیان شده است. آزمایش‌ها بر روی مجموعه‌داده‌های دارای دو طبقه انجام شده است. خصوصیات مجموعه‌داده ها در جدول (۱) نمایش داده شده است. به منظور ارزیابی عملکرد، روش پیشنهادی با [۲۰, ۳۲, ۳۳] مقایسه شده است. در داده‌های نامتوازن، طبقه کمینه از اهمیت بالایی برخوردار است. بدین منظور، معیارهای باز فراخوانی **Error! Reference source not found.** دقت F1 **Error! Reference source not found.** و G-mean **Error! Reference source not found.** مقایسه شده‌اند. اعتبارسنجی به شیوه LOO[1] انجام شده است.

(جدول -۱) مجموعه‌داده

(Table -1) Dataset

مجموعه‌داده	تعداد نمونه	تعداد ویژگی	تعداد طبقه
Sea10	۱۰۰۰۰	۳	۲
Weather	۱۰۰۰۰	۸	۲
Sea60	۶۰۰۰۰	۳	۲
Stagger	۲۰۰۰۰	۳	۲
Hyperplane	۱۰۰۰۰۰	۱۰	۲
Electricity	۴۵۳۱۲	۸	۲

معیار باز فراخوانی برای ارزیابی طبقه کمینه استفاده می‌شود. دقت کسری از نمونه‌های طبقه کمینه را نشان می‌دهد که به عنوان نمونه‌های طبقه کمینه به‌درستی طبقه‌بندی شده‌اند. F1 میانگین هارمونیک باز فراخوانی و دقت است. G-mean یا میانگین هندسی دقت دو طبقه بیشینه و کمینه را محاسبه نموده و سعی می‌کند ضمن به دست آوردن توازن مناسب، دقت دو طبقه را به بیشینه برساند.

$$\text{Recall} = \frac{TP}{TP+FN} \quad (۲۰)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (۲۱)$$

$$\text{F1} = \frac{2\text{Recall.Precision}}{\text{Recall+Precision}} \quad (۲۲)$$

$$\text{G-mean} = \sqrt{\frac{TN}{TN+FP} \times \frac{TP}{TP+FN}} \quad (۲۳)$$

که *TP*، تعداد داده‌هایی که به‌درستی مثبت (رانش مفهوم) تشخیص داده شده‌اند. *FN*، تعداد داده‌هایی که به‌اشتباه منفی تشخیص داده شده‌اند. *FP*، تعداد داده‌هایی که به‌اشتباه مثبت تشخیص داده شده‌اند. *TN*، تعداد داده‌هایی که به‌درستی منفی تشخیص داده شده‌اند.

۵-۱-نتایج

در این بخش نتایج حاصل از اجرای روش‌های مختلف به ازای عملکرد طبقه‌بندها، به ازای معیارهای باز فراخوانی، دقت، F1 و G-mean در جداول ۵-۲ آورده شده‌اند. برای سهولت در مقایسه، رتبه هر روش به‌ازای هر مجموعه‌داده و همچنین میانگین رتبه هر روش به صورت مجزا محاسبه و برترین نتایج به‌صورت پررنگ نشان داده شده‌اند. همان‌طور که مشاهده می‌شود روش پیشنهادی توانسته است در تمامی معیارها بهترین عملکرد را داشته باشد.

(جدول- ۲) باز فراخوانی

(Table -2) Recall

مجموعه‌داده	[۳۲]	[۳۳]	[۲۰]	روش پیشنهادی
Sea10	۹۵.۳۳ (۲)	۹۵.۳۰ (۳)	۹۵.۰۲ (۴)	**۹۵.۴۰ (۱)**
Weather	۸۲.۲۳ (۴)	۸۲.۲۳ (۳)	**۸۴.۴۸ (۱)**	۸۴.۱۳ (۲)
Sea60	۷۰.۸۴ (۴)	۷۵.۱۴ (۲)	۷۵.۰۲ (۳)	**۷۵.۶۸ (۱)**
Stagger	۹۶.۵۹ (۳)	۹۴.۱۸ (۴)	۹۷.۱۸ (۲)	**۹۷.۳۸ (۱)**
hyperplaneX	**۸۵.۶۴ (۱)**	۸۴.۸۱ (۳)	۸۵.۳۳ (۲)	۸۴.۵۶ (۴)
Electricity	**۹۰.۲۶**	۸۸.۹۵	۸۹.۹۲	۹۰.۱۴ (۲)

[1] Leave-One-Out cross-validation

	(۱)	(۴)	(۳)	
میانگین رتبه	(۳) ۲.۵	(۴) ۳.۱	(۲) ۲.۳	**(۱) ۱.۸**

(جدول-۳) دقت

(Table -3) Recall

مجموعه‌داده	**[۳۲]**	**[۳۳]**	**[۲۰]**	**روش پیشنهادی**
Sea10	**۹۰.۰۸** (۳)	**۸۹.۸۹** (۴)	**۹۰.۱۸** (۲)	**(۱) ۹۰.۲۴**
Weather	**۸۴.۹۶** (۲)	**۸۴.۶۶** (۳)	**۸۴.۳۴** (۴)	**(۱) ۸۵.۲۸**
Sea60	**۹۱.۲۰** (۴)	**۹۱.۲۹** (۳)	**۹۱.۷۸** (۲)	**(۱) ۹۱.۸۸**
Stagger	**۸۹.۴۱** (۴)	**۸۹.۵۹** (۳)	**۸۹.۹۳** (۱)	**(۲) ۸۹.۸۹**
hyperplaneX	**۸۸.۰۶** (۴)	**۸۸.۹۳** (۳)	**۹۰.۳۵** (۱)	**(۲) ۹۰.۲۹**
Electricity	**۷۵.۸۳** (۴)	**۷۸.۴۹** (۳)	**۷۹.۷۵** (۱)	**(۲) ۷۹.۳۹**
میانگین رتبه	**(۴) ۳.۵**	**(۳) ۳.۱**	**(۲) ۱.۶**	**(۱) ۱.۵**

(جدول -۴) F1

(Table- 4) F1

مجموعه‌داده	**[۳۲]**	**[۳۳]**	**[۲۰]**	**روش پیشنهادی**
Sea10	۹۲.۶۳ (۲)	(۴) ۹۲.۵۲	۹۲.۵۳ (۳)	**(۱) ۹۲.۷۵**
Weather	۸۳.۵۷ (۴)	(۳) ۸۳.۵۸	۸۴.۴۱ (۲)	**(۱) ۸۴.۷۰**
Sea60	۷۹.۸۴ (۴)	(۳) ۸۲.۴۳	۸۲.۵۶ (۲)	**(۱) ۸۲.۹۹**
Stagger	۹۲.۸۶ (۳)	(۴) ۹۱.۸۲	۹۳.۴۱ (۲)	**(۱) ۹۳.۴۹**
hyperplaneX	۸۶.۸۳ (۳)	(۴) ۸۶.۸۲	**۸۷.۷۲** (۱)	(۲) ۸۷.۳۳
Electricity	۸۲.۴۲ (۴)	(۳) ۸۳.۳۹	**۸۴.۵۳** (۱)	(۲) ۸۴.۴۲
میانگین رتبه	(۳) ۳.۱	(۴) ۳.۵	(۲) ۱.۸	**(۱) ۱.۳**

(جدول- ۵) G-mean

Table 5 G-mean

مجموعه‌داده	**[۳۲]**	**[۳۳]**	**[۲۰]**	**روش پیشنهادی**
Sea10	۸۷.۷۵ (۲)	۸۷.۶۲ (۴)	۸۷.۶۹ (۳)	**(۱) ۸۷.۹۳**
Weather	۷۳.۶۹ (۳)	۷۳.۷۲ (۲)	۷۳.۴۵ (۴)	**(۱) ۷۴.۶۳**
Sea60	۸۲.۳۱ (۴)	۸۴.۸۲ (۳)	۸۴.۸۳ (۲)	**(۱) ۸۵.۲۳**
Stagger	**۸۸.۱۸** (۱)	۸۶.۶۵ (۴)	۸۸.۱۷ (۲)	(۳) ۸۸.۱۲
hyperplaneX	۸۶.۷۹	۸۶.۹۰	**۸۷.۵۸**	(۲) ۸۷.۴۵

	(۱)	(۳)	(۴)	
Electricity	**۷۸.۴۰** (۱)	۷۵.۷۳ (۳)	۷۲.۳۹ (۴)	(۲) ۷۸.۲۹
میانگین رتبه	۲.۱۷ (۲)	۳.۱۷ (۴)	(۳) ۳.۰۰	(۱) ۱,۶۷

۶- نتیجه‌گیری و کارهای آینده

در این مقاله، با هدف حل مسئله طبقه‌بندی داده‌های جریانی نامتوازن دارای تغییر مفهوم ارائه شده است. با ورود داده جریانی، ابتدا داده‌های نویزی و داده‌های مرزی با انجام خوشه‌بندی شناسایی شده و تنها داده‌های مطمئن، با کمک تابع باور وزن دهی می‌شوند. با در نظر گرفتن برچسب داده‌ها، نمونه‌افزایی تنها در نواحی کم تراکم داده‌های طبقه کمینه و با رویکرد آشوبی انجام می‌گیرد. سپس، با استفاده از تعریف حد آستانه، رانش مفهوم شناسایی می‌شود. پیش‌بینی برچسب توسط طبقه‌بند ترکیبی و رای گیری وزن‌دار بیشینه انجام می‌پذیرد. روش پیشنهادی با تلفیق یادگیری جمعی و یادگیری افزایشی از سه مزیت بهره می‌برد. (۱) با ایجاد طبقه‌بندهای کاندید و رویکرد وزن دهی، نمونه‌هایی که نیازمند آموزش بیشتری هستند شناسایی شده و موردتوجه بیشتری قرار می‌گیرند. (۲) مسئله عدم توازن در توزیع طبقات، با نمونه‌افزایی در نواحی کم تراکم طبقه کمینه حل می‌شود. (۳) با تعریف سه حد آستانه، وقوع یا عدم وقوع رانش مفهوم تشخیص داده شده و با رأی‌گیری وزن‌دار بیشینه، برچسب نمونه‌ها پیش‌بینی می‌شود؛ بنابراین رانش مفهوم در هر قطعه تشخیص داده شده و داده فاقد رانش برای آموزش طبقه‌بندها مورداستفاده قرار می‌گیرد. نکته مهم آن است به دلیل آنکه مقادیر حد آستانه باید توسط کاربر و پیش از اجرا تعیین شوند ممکن است در آغاز کار به‌عنوان نقطه‌ضعف شناخته شوند. علاوهبر این، پیچیدگی ناشی از انتخاب نزدیک‌ترین k همسایه بر سرعت طبقه‌بند تأثیرگذار است. عملکرد روش پیشنهادی بر روی داده‌های برگرفته‌شده از پایگاه داده UCI توسط روش LOO ارزیابی و با طبقه‌بندهای مرز دانش مقایسه شده است. آزمایش‌ها نشان‌دهنده برتری روش پیشنهادی از نظر معیارهای باز فراخوانی، دقت، F1 و G-mean است.

جهت کارهای آینده پیشنهاد می‌شود که با ترکیب روش پیشنهادی با سایر روش‌های یادگیری نظارتی، روش‌های یادگیری فعال کارامدی برای داده‌های جریانی ارائه شود.

[DOI: 10.61186/jsdp.20.4.23]

7-Refrence ۷- مراجع

[1] G. Douzas, R. Rauch, and F. Bacao, "G-SOMO: An oversampling approach based on self-organized maps and geometric SMOTE," *Expert Systems with Applications*, vol. 183, p. 115230, 2021, doi: https://doi.org/10.1016/j.eswa.2021.115230.

[2] J. Engelmann and S. Lessmann, "Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning," *Expert Systems with Applications*, vol. 174, p. 114582, 2021, doi: https://doi.org/10.1016/j.eswa.2021.114582.

[3] X. Xie, H. Liu, S. Zeng, L. Lin, and W. Li, "A novel progressively undersampling method based on the density peaks sequence for imbalanced data," *Knowledge-Based Systems*, vol. 213, p. 106689, 2021, doi: https://doi.org/10.1016/j.knosys.2020.106689.

[4] G. Douzas, F. Bacao, and F. Last, "Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE," *Information Sciences*, vol. 465, pp. 1-20, 2018.

[5] Z. Xu, D. Shen, T. Nie, Y. Kou, N. Yin, and X. Han, "A cluster-based oversampling algorithm combining SMOTE and k-means for imbalanced medical data," *Information Sciences*, vol. 572, pp. 574-589, 2021, doi: https://doi.org/10.1016/j.ins.2021.02.056.

[6] Z. Li, W. Huang, Y. Xiong, S. Ren, and T. Zhu, "Incremental learning imbalanced data streams with concept drift: The dynamic updated ensemble algorithm," *Knowledge-Based Systems*, vol. 195, p. 105694, 2020, doi: 10.1016/j.knosys.2020.105694.

[7] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1/2, pp. 100-115, 1954.

[8] D. Siegmund, *Sequential analysis: tests and confidence intervals*. Springer Science & Business Media, 2013.

[9] O. A. Mahdi, E. Pardede, and N. Ali, "KAPPA as Drift Detector in Data Stream Mining," *Procedia Computer Science*, vol. 184, pp. 314-321, 2021.

[10] J. Gama, P. Medas, G. Castillo, and P. Rodrigues, "Learning with Drift Detection," Berlin, Heidelberg, 2004: Springer Berlin Heidelberg, in Advances in Artificial Intelligence – SBIA 2004, pp. 286-295.

[11] M. Baena-Garcıa, J. del Campo-Ávila, R. Fidalgo, A. Bifet, R. Gavalda, and R. Morales-Bueno, "Early drift detection method," in *Fourth international workshop on knowledge discovery from data streams*, 2006, vol. 6, pp. 77-86.

[12] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*, 2007: SIAM, pp. 443-448.

[13] P. Domingos and G. Hulten, "Mining high-speed data streams," in *Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, 2000: ACM, pp. 71-80.

[14] G. Liu, H. Cheng, Z. Qin, Q. Liu, and C. Liu, "E-CVFDT: An improving CVFDT method for concept drift data stream," in *2013 International Conference on Communications, Circuits and Systems (ICCCAS)*, 2013, vol. 1, pp. 315-318, doi: 10.1109/ICCCAS.2013.6765241.

[15] S. A. Jadhav and S. Kosbatwar, "Concept-adapting Very Fast Decision Tree with Misclassification Error," 2016.

[16] I. Frias-Blanco, J. del Campo-Ávila, G. Ramos-Jimenez, R. Morales-Bueno, A. Ortiz-Diaz, and Y. Caballero-Mota, "Online and non-parametric drift detection methods based on Hoeffding's bounds," *IEEE Transactions on Knowledge Data Engineering*, vol. 27, no. 3, pp. 810-823, 2014.

[17] A. Pesaranghader and H. L. Viktor, "Fast hoeffding drift detection method for evolving data streams," in *Joint European conference on machine learning and knowledge discovery in databases*, 2016: Springer, pp. 96-111.

[18] A. Pesaranghader, H. Viktor, and E. Paquet, "Reservoir of diverse adaptive learners and stacking fast hoeffding drift detection methods for evolving data streams," *Machine Learning*, vol. 107, no. 11, pp. 1711-1743, 2018.

[19] Y. Yuan, Z. Wang, and W. Wang, "Unsupervised concept drift detection based on multi-scale slide windows," *Ad Hoc Networks*, vol. 111, p. 102325, 2021.

[20] A. Feitosa Neto and A. M. P. Canuto, "EOCD: An ensemble optimization approach for concept drift applications," *Information Sciences*, vol. 561, pp. 81-100, 2021, doi: https://doi.org/10.1016/j.ins.2021.01.051.

[21] D. H. Jeong and J. M. Lee, "Ensemble learning based latent variable model predictive control for batch trajectory tracking under concept drift," *Computers & Chemical Engineering*, vol. 139, p. 106875, 2020, doi: https://doi.org/10.1016/j.compchemeng.2020.106875.

[22] W. Liu, H. Zhang, Z. Ding, Q. Liu, and C. Zhu, "A comprehensive active learning method for multiclass imbalanced data streams with concept drift," *Knowledge-Based Systems*, vol. 215, p. 106778, 2021, doi: https://doi.org/10.1016/j.knosys.2021.106778.

[23] P. Zyblewski, R. Sabourin, and M. Woźniak, "Preprocessed dynamic classifier ensemble selection for highly imbalanced drifted data streams," *Information Fusion*, vol. 66, pp. 138-154, 2021, doi: 10.1016/j.inffus.2020.09.004.

[24] J. Hamidzadeh and M. Moradi, "Improving Chernoff criterion for classification by using the filled function," *jsdp*, vol. 19, no. 3, pp. 105-118, 2022, doi: 10.52547/jsdp.19.3.105.

[25] J. Pouramini, B. Minaei-Bidgoli, and M. Esmaeili, "A Novel One Sided Feature Selection Method for Imbalanced Text Classification," *jsdp*, vol. 16, no. 1, pp. 21-40, 2019, doi: 10.29252/jsdp.16.1.21.

[26] E. Yasrebi Naeini and m. hatami, "Improving Imbalanced Data Classification Accuracy by using Fuzzy Similarity Measure and Subtractive Clustering," *jsdp*, vol. 19, no. 2, pp. 27-38, 2022, doi: 10.52547/jsdp.19.2.27.

[27] Y. Wang, Y. Zhang, and Y. Wang, "Mining Data Streams with Skewed Distribution by Static Classifier Ensemble," in *Opportunities and Challenges for Next-Generation Applied Intelligence*, B.-C. Chien and T.-P. Hong Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 65-71.

[28] S. Chen and H. He, "Towards incremental learning of nonstationary imbalanced data stream: a multiple selectively recursive approach," *Evolving*

[DOI: 10.61186/jsdp.20.4.23]

Systems, vol. 2, no. 1, pp. 35-50, 2011, doi: 10.1007/s12530-010-9021-y.
[29] R. N. Lichtenwalter and N. V. Chawla, "Adaptive Methods for Classification in Arbitrarily Imbalanced and Drifting Data Streams," in *New Frontiers in Applied Data Mining*, Berlin, Heidelberg, T. Theeramunkong *et al.*, Eds., 2010// 2010: Springer Berlin Heidelberg, pp. 53-75.
[30] G. Ditzler and R. Polikar, "Incremental Learning of Concept Drift from Streaming Imbalanced Data," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 10, pp. 2283-2301, 2013, doi: 10.1109/TKDE.2012.136.
[31] R. R. Yager and L. Liu, *Classic works of the Dempster-Shafer theory of belief functions*. Springer, 2008.
[32] M. A. A. Abdualrhman and M. Padma, "CD2A: Concept Drift Detection Approach Toward Imbalanced Data Stream," in *Emerging Research in Electronics, Computer Science and Technology*: Springer, 2019, pp. 597-612.
[33] M. M. W. Yan, "Accurate detecting concept drift in evolving data streams," *ICT Express*, 2020.

نشانی رایانامه ایشان عبارت است از:
MonaMoradi0@gmail.com

جواد حمیدزاده در حال حاضر دانشیار دانشکده مهندسی کامپیوتر و فناوری اطلاعات دانشگاه سجاد است. از علاقه‌مندی‌های ایشان می‌توان به یادگیری ماشین، رایانش نرم، بازشناسی الگو و شبکه‌های رایانه‌ای اشاره کرد.
نشانی رایانامه ایشان عبارت است از:
J_hamidzadeh@sadjad.ac.ir

محمدعلی رشیدی محمودی دارای مدرک کارشناسی ارشد مهندسی رایانه در گرایش نرم‌افزار است. از علاقه‌مندی‌های ایشان می‌توان به بازشناسی الگو، داده حجیم و یادگیری عمیق اشاره کرد.
نشانی رایانامه ایشان عبارت است از:
Ma.rashidi191@sadjad.ac.ir

منا مرادی دارای مدرک کارشناسی ارشد مهندسی رایانه در گرایش نرم‌افزار است. از علاقه‌مندی‌های ایشان می‌توان به رایانش نرم، یادگیری ماشین و بازشناسی الگو اشاره کرد.

[DOI: 10.61186/jsdp.20.4.23]