

تشخیص تغییر مفهوم در جریان داده با کمک

رده‌بند نیمه‌نظارتی

حسین حسن‌نژاد نامقی^۱، هدی مشایخی^{*} و مرتضی زاهدی^۱
^۱دانشکده مهندسی کامپیوتر، دانشگاه صنعتی شاهرود، شاهرود، ایران

چکیده

جریان داده به دنباله‌ای از داده‌ها گفته می‌شود که از منابع اطلاعاتی مختلف با سرعت زیاد و حجم بالا تولید می‌شوند. از مهم‌ترین چالش‌های موجود در تحلیل جریان داده وجود تغییر مفهوم در آن‌ها است. تغییر مفهوم به معنای تغییر ویژگی‌های آماری داده‌هاست. در بسیاری از پژوهش‌های موجود برای مقابله با چالش نامحدود بودن طول جریان داده و یا چالش تغییر مفهوم، از رویکردهایی با فرض موجود بودن برچسب درست برای همه داده‌ها استفاده می‌کنند؛ در حالی که با توجه به هزینه‌بر بودن فرآیند برچسب‌دهی جریان داده، به‌طور عمومی فرض می‌شود تنها بخشی از داده‌ها دارای برچسب هستند. در این مقاله یک روش یادگیری گروهی نیمه‌نظارتی ارائه شده که از تغییر آنتروپی برای تشخیص تغییر مفاهیم در رده‌بندی جریان داده استفاده می‌کند. مدل یادگیری گروهی پیشنهادی با تعداد محدودی داده برچسب‌دار اولیه آموزش می‌بیند؛ سپس در صورت مشاهده تغییر مفهوم، از داده‌های بدون برچسب برای به‌روزرسانی مدل رده‌بند گروهی استفاده می‌کند. روش پیشنهادی قادر است تغییرات موجود در مجموعه داده را تشخیص داده و با به‌روزرسانی مدل یادگیری، در بهبود دقت الگوریتم مؤثر باشد. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی از جنبه‌های مختلف نسبت به سایر روش‌ها کارایی بالاتری دارد.

واژگان کلیدی: جریان داده، یادگیری گروهی، تغییر مفهوم، آنتروپی، رده‌بند نیمه‌نظارتی

Detecting Concept Drift in Data Stream Using Semi-Supervised Classification

Hossein Hasan Nezhad Namaghi¹, Hoda Mashayekhi^{*1} & Morteza Zahedi¹

¹Faculty of Computer Engineering, Shahrood University of Technology, Shahrood, Iran

Abstract

Data stream is a sequence of data generated from various information sources at a high speed and high volume. Classifying data streams faces the three challenges of unlimited length, online processing, and concept drift. In related research, to meet the challenge of unlimited stream length, commonly the stream is divided into fixed size windows or gradual forgetting is used. Concept drift refers to changes in the statistical properties of data, and is divided into four categories: sudden, gradual, incremental, and recurring. Concept drift is generally dealt with by periodically updating the classifier, or employing an explicit change detector to determine the update time. These approaches are based on the assumption that the true labels are available for all data samples. Nevertheless, due to the cost of labeling instances, access to a partial labeling is more realistic. In a number of studies that have used semi-supervisory learning, the labels are received from the user to update the models in form of active learning. The purpose of this study is to classify samples in an unlimited data stream in presence of concept drift, using only a limited set of initial labeled data. To this end, a semi-supervised ensemble learning algorithm for data stream is proposed, which uses entropy variation to detect concept drift and is applicable for sudden and gradual drifts. The proposed model is trained with a limited initial labeled set. In occurrence of concept drift, the unlabeled data is used to update the ensemble model. It does not

* Corresponding author

* نویسنده عهده‌دار مکاتبات

سال ۱۴۰۰ شماره ۴ پیاپی ۵۰

• تاریخ ارسال مقاله: ۱۳۹۸/۰۳/۱۸ • تاریخ پذیرش: ۱۳۹۹/۱۲/۱۱ • تاریخ انتشار: ۱۴۰۰/۱۲/۲۹ • نوع مطالعه: پژوهشی



require receiving the labels from the user. In contrast to many of the current studies, the proposed algorithm uses an ensemble of K-NN classifiers. It constructs a group of clustering-based classification models, each of which is trained on a batch of data. On receiving each new sample, first it is determined whether the data sample is an outlier or not. If the data is included in a cluster, the sample class is determined by majority voting. When a window of the stream is received, the possibility of concept drift is examined based on entropy variation, and the classifier is updated by a semi-supervised approach if necessary. The model itself determines the required data labels. The proposed method is capable of detecting concept drift in data, and improving its accuracy via updating the learning model with appropriate samples received from the stream. Therefore, the proposed method only requires a small initial labeled data. Experiments are performed using five real and synthetic datasets, and the model performance is compared to three other approaches. The results show that the proposed method is superior in terms of precision, recall and F1 score compared to other studies.

Keywords: data stream, ensemble learning, concept drift, entropy, semi-supervised classification

۱- مقدمه

پیشرفت‌های اخیر سخت‌افزاری در زمینه‌های ذخیره‌سازی و پردازش اطلاعات، امکان جمع‌آوری خودکار داده‌ها را فراهم کرده است. به این گونه داده‌ها که به سرعت در حال تولید و افزایش بوده و نیاز به تجزیه و تحلیل آنی در هنگام ورود دارند، جریان داده^۱ گفته می‌شود. برای نمونه می‌توان به تراکنش‌های مالی، داده‌های موجود در شبکه‌های اجتماعی و داده‌های به دست آمده از حسگرهای بی‌سیم اشاره کرد [1].

در کاوش جریان‌های داده از روش‌های گوناگونی هم‌چون رده‌بندی^۲، خوشه‌بندی^۳، درخت تصمیم^۴ و شبکه‌های عصبی^۵ استفاده می‌شود؛ ولی در این مقاله تمرکز بیشتر بر روی رده‌بندی است. کاوش جریان‌های داده با سه چالش طول نامحدود، پردازش برخط و تغییر مفهوم مواجه است [2]. استفاده از الگوریتم‌های متداول یادگیری به دلیل نامحدود بودن جریان داده، عدم توانایی ذخیره‌سازی کل داده‌ها و همچنین زمان یادگیری نامحدود آن‌ها، میسر نیست. فرآیند یادگیری در این داده‌ها باید با یک گذر روی داده‌ها انجام شود، اما الگوریتم‌های متداول داده‌کاوی نیاز به چندین گذر روی داده‌ها و دسترسی به داده‌های قدیمی دارند، که به دلیل محدودیت حافظه کارایی لازم را ندارند و از سوی دیگر، جهت پردازش حجم عظیم جریان‌های داده بسیار کند و غیر عملی هستند [11، 28]. تغییر مفهوم^۶ زمانی اتفاق می‌افتد که توزیع داده‌ها در طول زمان به طور اساسی تغییر کند [12]؛ بنابراین مدل رده‌بند باید با توجه به توزیع جدید به صورت پیوسته به روز شود.

در پژوهش‌های مرتبط، برای مقابله با چالش نامحدود بودن طول جریان داده، از تقسیم جریان داده به قطعاتی با اندازه ثابت [3]، یا فراموشی تدریجی^۷ [4] استفاده شده است. برای مواجهه با چالش تغییر مفهوم از به روزرسانی دوره‌ای رده‌بند [1]، یا تشخیص‌دهنده تغییر صریح [5] برای تشخیص زمان به روزرسانی استفاده شده است. این رویکردها با فرض اینکه برچسب‌های درست برای تمام نمونه‌های داده موجود است، مطرح شده‌اند. درحالی‌که به طور عمومی فرآیند برچسب‌گذاری داده‌ها یا توسط یک کاربر به صورت دستی ارائه شده و یا اینکه نیازمند وسایل خاص و آزمایش‌های کند و گران است. با توجه به محدودیت زمان و منبع، امکان اینکه برچسب درست برای همه نمونه‌ها در دسترس باشد، وجود ندارد [6]؛ بنابراین باید به دنبال راهی بود تا از داده‌های بدون برچسب به عنوان داده‌های آموزشی استفاده کرد.

روش‌های یادگیری را می‌توان با توجه به میزان داده‌های آموزشی برچسب‌دار به سه دسته با نظارت^۸، نیمه‌نظارتی^۹ و بدون نظارت^{۱۰} رده‌بندی کرد. اگر همه داده‌های آموزشی برچسب‌دار باشند، یادگیری با نظارت خواهد بود. اگر هیچ یک از داده‌ها برچسب‌گذاری نشده باشند، یادگیری بدون نظارت است و در صورتی که علاوه بر داده‌های برچسب‌دار، از داده‌های بدون برچسب در یادگیری استفاده شود، یادگیری نیمه‌نظارتی خواهد بود. در روش یادگیری نیمه‌نظارتی، تعداد داده‌های برچسب‌دار در مقایسه با داده‌های بدون برچسب بسیار اندک است [11]. روش‌های گوناگونی برای رده‌بندی جریان‌های داده ارائه شده است، ولی به طور کلی الگوریتم‌های رده‌بندی با در نظر گرفتن تغییر مفهوم به دو دسته الگوریتم‌های کور و

⁷ Gradual Forgetting

⁸ Supervised

⁹ Semi-Supervised

¹⁰ Unsupervised

¹ Data Stream

² Classification

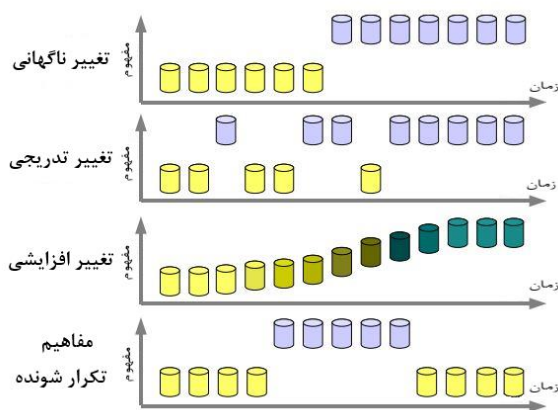
³ Clustering

⁴ Decision Tree

⁵ Neural Networks

⁶ Concept Drift

قابلیت‌های جدید به آن، اشاره کرد.
۴. مفهوم تکرارشونده: در این نوع تغییر مفهوم، مفهومی که در گذشته فعال بوده، می‌تواند بعد از مدت زمانی دوباره پدیدار شود. برای مثال در مسأله جستجوی خودرو ممکن است، کاربر دوباره قصد خرید خودرو را داشته باشد.



(شکل-۱): انواع تغییر مفهوم [13]
(Figure-1): The kinds of concept drift [13]

در برخی از پژوهش‌های انجام‌شده در زمینه تشخیص تغییر، روش‌هایی ارائه شده‌اند که به برچسب درست برای تمام نمونه‌ها نیاز دارند و درحقیقت از روش یادگیری نظارتی برای شناسایی تغییر استفاده می‌کنند. در تعدادی از پژوهش‌های انجام‌گرفته که از یادگیری نیمه‌نظارتی استفاده کرده‌اند، برای به‌روزرسانی مدل‌ها برچسب نمونه‌ها را از کاربر دریافت کرده و از روش یادگیری فعال استفاده می‌کنند. در این مقاله روشی برای رده‌بندی تدریجی جریان داده مبتنی بر مدل‌های گروهی ارائه شده که از آنتروپی به‌عنوان شناسایی‌کننده تغییر استفاده کرده و دو نوع تغییر مفهوم تدریجی و ناگهانی را شناسایی می‌کند. روش یادگیری به‌صورت نیمه‌نظارتی است و از داده‌های آموزشی محدودی برای ساخت مدل‌های اولیه استفاده می‌کند. روش پیشنهادی برای به‌روزرسانی مدل‌ها هیچ برچسب نمونه‌ای را از کاربر دریافت نمی‌کند.

قالب ارائه‌شده دارای گروهی از مدل‌های رده‌بند مبتنی بر خوشه‌بندی است که هر کدام روی یک قطعه^۷ از داده آموزش می‌بیند؛ سپس با ورود هر نمونه آزمایش، مشخص می‌شود نمونه داده پرت^۸ است یا خیر. در صورتی که داده درون خوشه قرار گرفت، با استفاده از رأی اکثریت رده نمونه تعیین می‌شود و با استفاده از مدل

آگاه تقسیم می‌شوند. در روش‌های کور بدون توجه به تغییر مفهوم، مدل در زمان‌هایی مشخص به‌روز می‌شود؛ ولی در روش‌های آگاه با بررسی جریان داده، در صورت رخداد تغییر مفهوم مدل به‌روزرسانی خواهد شد [13].

همان‌گونه که بیان شد، تغییر مفهوم تغییر توزیع ویژگی یا برچسب نمونه‌هاست. اگر تغییر مفهوم در توزیع داده‌ها (ویژگی‌ها) رخ دهد، تغییر مفهوم مجازی^۱ و اگر در توزیع برچسب داده‌ها باشد، تغییر مفهوم واقعی^۲ نامیده می‌شود [12]. تغییر مفهوم به چهار دسته ناگهانی^۳، تدریجی^۴، افزایشی^۵ و تکرارشونده^۶ تقسیم می‌شود. با فرض این که فقط دو مفهوم داریم، این چهار نوع تغییر را بررسی می‌کنیم [13]:

۱. تغییر مفهوم ناگهانی: ساده‌ترین نوع تغییر مفهوم است، به‌طوری‌که در یک زمان مشخص مفهوم جدیدی جایگزین مفهوم قبلی می‌شود. برای نمونه شخصی قصد خرید خودرو دارد و برای بررسی اطلاعات خودرو و قیمت آن در اینترنت جستجو می‌کند؛ بنابراین بسیاری از جستجوها در ارتباط با خودرو است. درحالی‌که پس از خرید خودرو، جستجوها تغییر می‌کنند.

۲. تغییر مفهوم تدریجی: در این نوع تغییر مفهوم، مفهوم به‌تدریج از یک نوع به نوعی دیگر تغییر می‌کند. در یک بازه زمانی هر دو مفهوم فعال هستند؛ ولی به‌تدریج از احتمال انتخاب نمونه‌ها برای نوع اول کم شده و به‌احتمال انتخاب نمونه‌ها برای نوع دوم اضافه می‌شود. برای نمونه زمانی که سرمربی تیم یک کشور تغییر می‌کند، در آن کشور جستجوها در ارتباط با آن شخص به‌مرور کم شده و برای شخصی که به‌تازگی سرمربی تیم شده به‌تدریج زیاد می‌شود.

۳. تغییر مفهوم افزایشی: این نوع تغییر مفهوم شباهت زیادی به تغییر مفهوم تدریجی دارد، با این تفاوت که بیش از دو نوع مفهوم وجود دارد که بین این دو نوع مفهوم اصلی قرار می‌گیرند. در یک بازه زمانی توزیع نمونه‌ها به‌تدریج از نوع اول فاصله گرفته و به نوع دوم نزدیک می‌شود. برای مثال می‌توان به افزایش تدریجی محبوبیت یک نرم‌افزار پس از اضافه‌شدن

¹ Virtual

² Real

³ Sudden

⁴ Gradual

⁵ Incremental

⁶ Recurring

⁷ Chunk

⁸ Outlier

ارائه شده، امکان وجود تغییر مفهوم بررسی می شود؛ در نهایت در صورت لزوم رده بند به روزرسانی می شود.

روش پیشنهادی با برخی از روش های موجود بر روی مجموعه داده های مختلف مقایسه شده است. نتایج آزمایش ها نمایان گر عملکرد بهتر روش پیشنهادی نسبت به سایر روش ها است. به طوری که این روش به خوبی توانسته تغییرات موجود در مجموعه داده را تشخیص داده و دقت و صحت الگوریتم را بهبود بخشد. در ادامه، در بخش ۲ مروری بر پژوهش های رده بندی جریان داده بررسی و در بخش ۳ چارچوب پیشنهادی برای رده بندی جریان داده با تشخیص تغییر مفهوم ارائه می شود. در بخش ۴ به بررسی نتایج و مقایسه آن ها می پردازیم و در پایان نتیجه گیری ارائه می شود.

۲- کارهای انجام شده

در یادگیری دسته ای^۱ با میانگین گیری از مدل های تصادفی حاصل از آموزش داده ها، بر مشکل محدود بودن داده های آموزشی غلبه می شود [14]. اما این روش به حجم داده ها وابسته بوده و با افزایش حجم داده ها از نظر زمان اجرا برای تولید مدل ها و تحلیل و بررسی آن ها دچار مشکل می شود. امروزه بیش تر پژوهش های رده بندی جریان داده با هدف رسیدن به کارایی مطلوب و به کمینه رساندن خطا انجام می شوند.

روش های تشخیص تغییر مفهوم به صورت کلی به دو دسته مبتنی بر کارایی و مبتنی بر توزیع داده تقسیم می شوند [27]. روش های مبتنی بر کارایی به صورت پیوسته دنباله مقادیر معیارهای کارایی مثل دقت را رصد می کنند و کاهش عمده در این معیارها نشان دهنده تغییر مفهوم خواهد بود. این روش ها نیاز به برچسب داده ها داشته و با ناظر هستند. در این دسته برخی از مقالات از وزن دهی برای تشخیص تغییر مفهوم استفاده می کنند. برای نمونه سیدو و همکارش [8] یک روش مبتنی بر وزن دهی پویا برای رده بندی داده ها ارائه می دهد. این وزن دهی بر پایه دو معیار دقت رده بند و پیش بینی نهایی الگوریتم است. کابرال و همکاران [18] روشی مبتنی بر آزمایش فیشر پیشنهاد داده و سه کاربرد آن را در تشخیص تغییر مفهوم بررسی کرده اند. از آنجا که الگوریتم ارائه شده در این مقاله در دسته تشخیص تغییر مبتنی بر توزیع داده قرار می گیرد، در ادامه تمرکز بیشتری بر روش های این گروه قرار می دهیم.

در دسته دوم برخی از پژوهش ها از روش تشخیص تغییر صریح^۲ برای شناسایی تغییر مفهوم در جریان داده استفاده می کنند. در کاوش جریان داده، این روش یا برای تشخیص تغییر در توزیع داده ها و یا برای تشخیص تغییر در بازخورد رده بند استفاده می شود. برای نمونه کنچو و همکارش [7] تغییر توزیع داده ورودی در یک جریان داده را تشخیص می دهد. در پژوهشی دیگر بیفت و همکارش [5] با استفاده از یک روش مبتنی بر پنجره کشویی، اندازه پنجره را با توجه به میزان تغییرات موجود در داده های پنجره مشخص می کند. در این روش زمانی یک تغییر تشخیص داده می شود که میزان خطا در کل پنجره جاری به طور قابل ملاحظه ای از کمترین میزان خطای ثبت شده، بیشتر باشد. مشکل روش های بالا آن است که برای شناسایی تغییرات نیاز به برچسب درست برای تمام نمونه ها دارند.

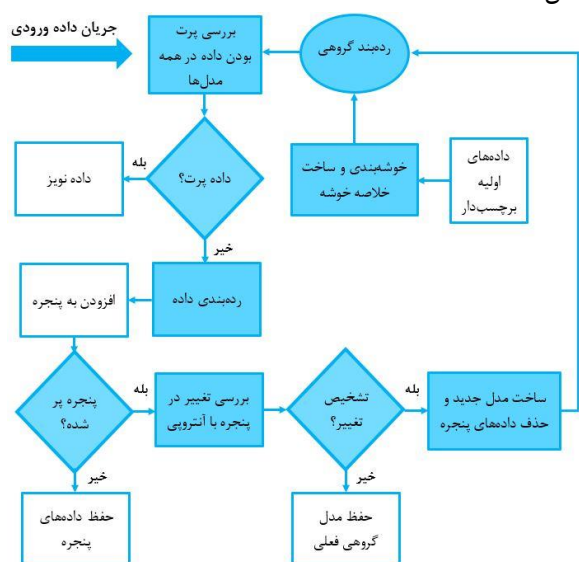
هک و همکارانش [6] نیز از یک روش تشخیص صریح برای تغییر مفهوم استفاده می کند، با این تفاوت که تشخیص تغییر در اطمینان رده بند انجام می گیرد و با توجه به این که تشخیص بر روی میزان خطای رده بند نیست، نیازی به برچسب درست برای تمام داده ها نداشته و از یک روش نیمه نظارتی بهره می برد. با ورود هر نمونه از داده های جریان، همراه با پیش بینی برچسب نمونه، میزان اطمینان پیش بینی صورت گرفته محاسبه و پس از اضافه شدن مقدار اطمینان به یک بافر، تشخیص تغییر با توجه به تغییرات توزیع مقادیر موجود در بافر بررسی می شود. برای به روزرسانی مدل ها از روش یادگیری فعال استفاده کرده و تعدادی برچسب از نمونه ها درخواست می کند.

استفاده از آنتروپی اطلاعات برای تشخیص تغییر مفهوم نخستین بار توسط وربرگر و برنستین [24] پیشنهاد گردید. دو و همکاران [25] این روش را با استفاده از یک پنجره کشویی پویا بهبود دادند. روش آنها تلاش دارد زمان مناسب برای آموزش دوباره رده بند را تشخیص دهد. مهدی و همکارانش [9] از روشی مبتنی بر آنتروپی برای تشخیص تغییر مفهوم استفاده کرده و از یک روش نیمه نظارتی برای یادگیری الگوریتم بهره می برند. روش یاد شده ابتدا دو پنجره با طول ثابت در نظر گرفته و برای هر یک از داده های جریان، بر مبنای احتمال وقوع نمونه مقادیری به دست می آورد و به پنجره اضافه می کند؛ پس از پر شدن پنجره ها، آنتروپی هر کدام از پنجره ها محاسبه و

² Explicit Concept Drift Technique

¹ Batch

در این پژوهش از یک رده‌بند گروهی نیمه‌نظارتی شامل مدل‌های K-NN استفاده کرده‌ایم. گروهی بودن رده‌بند موجب می‌شود بتوان علاوه بر داده‌های اولیه، از داده‌های جریان نیز برای رده‌بندی بهره برد و به‌مرور از داده‌های جدیدتر برای مدل‌سازی استفاده کرد. بدین ترتیب با توجه به این که مدل پیوسته با داده‌های جدید به‌روزرسانی می‌شود، دقت رده‌بندی بالا می‌رود. این روش نیازی به برچسب برای تمام داده‌ها نداشته و فقط به مجموعه کوچکی از داده‌های برچسب‌دار اولیه نیاز دارد. همچنین برای به‌روزرسانی رده‌بند، هیچ برچسب نمونه‌ای درخواست نمی‌کند؛ بنابراین روش ارائه‌شده برای مجموعه داده‌هایی که تعداد کمی از نمونه‌های آن برچسب درست در اختیار دارند، مناسب است. روش پیشنهادی برای تشخیص تغییر از روشی مبتنی بر آنتروپی استفاده می‌کند. این تشخیص تغییر موجب می‌شود در صورتی که در توزیع داده‌ها تغییر قابل توجهی اتفاق بیافتد مدل با به‌روزرسانی، خود را با این تغییر تطبیق داده و در بهبود دقت رده‌بندی مؤثر عمل کند.



(شکل-۲): شمای کلی روش پیشنهادی

(Figure-2): High level work flow of the proposed method

در شکل (۲) شمای کلی الگوریتم پیشنهادی نشان داده شده است. در ابتدا رده‌بند گروهی شامل مدل‌هایی است که روی داده‌های آموزشی اولیه آموزش دیده‌اند؛ سپس با ورود داده‌های جریان و در نظر گرفتن دو پنجره، احتمال تعلق هر کدام از این داده‌ها به‌ازای هر کدام از مدل‌ها محاسبه و میانگین آن‌ها به دو پنجره وارد می‌شوند؛ سپس اختلاف آنتروپی دو پنجره به‌دست آمده و در صورتی که اختلاف آن‌ها از حد آستانه‌ای بیشتر بود، تغییر مفهوم شناسایی می‌شود. زمانی که تغییر تشخیص

نمونه‌های دو پنجره با هم ترکیب می‌شوند و آنتروپی مشترک^۱ آن‌ها به‌دست می‌آید؛ سپس در صورتی که مقدار به‌دست آمده از حد آستانه‌ای^۲ کمتر باشد، تغییر مفهوم تشخیص داده می‌شود. در این روش جزئیات رده‌بند و همچنین چگونگی به‌دست آوردن احتمال وقوع هر نمونه برای اضافه‌شدن به پنجره مشخص نشده است. در رویکردی مشابه هاگ و کاسنکی [26] روش ERICS را برای تشخیص تغییر مفهوم پیشنهاد می‌کنند که در آن از تغییر آنتروپی و معیار KL مربوط به توزیع پارامترهای مدل، استفاده می‌شود.

دمیلو و همکاران [19] به بررسی تشخیص تغییر به‌صورت بدون ناظر پرداخته و معیارهای فاصله مختلف در جریان داده‌ها را برای تشخیص تغییر بررسی کرده‌اند. ونگ و همکاران [20] روشی را برای تشخیص تغییر مطرح می‌کنند که در دو مقیاس مختلف کار می‌کند. ابتدا با بررسی داده‌ها در سطح ریزدانگی بالاتر، به تشخیص تغییر می‌پردازد. در صورت تشخیص تغییر، داده‌ها در ریزدانگی پایین‌تر بررسی و محل تغییر مشخص می‌شود. برای تشخیص تغییر از نمونه‌گیری و آزمایش t-student استفاده می‌شود. سانگ و همکاران [21] از روش رگرسیون مبتنی خوشه‌بندی فازی برای تشخیص الگوها به‌صورت پویا و درجه تعلق هر نمونه به الگوهای شناسایی‌شده، استفاده کرده‌اند. لی و همکاران [22] یک روش نیمه‌نظارتی افزایشی مبتنی بر یادگیری عمیق ارائه می‌کنند. برای یادگیری ویژگی از اتوانکدر استفاده کرده و برای بهبود شبکه از محدودیت‌های شباهت و عدم شباهت جفت داده‌ها استفاده می‌کند. مدل آن‌ها در مقایسه با روش‌های بررسی‌شده قبلی پیچیده‌تر بوده و پارامترهای زیادی جهت آموزش دارد.

همان‌طور که در قبل اشاره شد روش ارائه‌شده در این مقاله در دسته روش‌های مبتنی بر توزیع داده قرار دارد و نیاز به اطلاع از برچسب تمامی داده‌ها ندارد. متفاوت با بسیاری از روش‌های یادشده، در این پژوهش برای افزایش دقت از رده‌بند گروهی استفاده کرده و برای تشخیص تغییر از روشی مبتنی بر آنتروپی بهره برده‌ایم. زمانی که تغییر تشخیص داده شود، بدون نیاز به درخواست برچسب برای نمونه‌های در جریان، از روشی نیمه‌نظارتی برای به‌روزرسانی رده‌بند استفاده می‌شود.

۳- روش پیشنهادی

¹ Joint Entropy

² Threshold

داده شود، مدل گروهی به روزرسانی و داده‌های دو پنجره حذف می‌شود.

۳-۱- آموزش و رده‌بندی

مجموعه داده‌ای که برای رده‌بندی به الگوریتم داده می‌شود، شامل مجموعه داده محدود اولیه برچسب‌دار که برای آموزش استفاده می‌شود و جریان داده ورودی که با طول نامحدود وارد می‌شود. به طوری که تعداد داده‌های آموزشی اولیه نسبت به تعداد داده‌های جریان کم بوده و فقط به برچسب داده‌های آموزشی برای اجرا بر روی الگوریتم نیاز داریم.

رده‌بند گروهی شامل مدل‌هایی است که با استفاده از الگوریتم K-NN روی داده‌های آموزشی اولیه، آموزش دیده‌اند. برای ساخت هر مدل، تعدادی از این نمونه‌ها را که به آن قطعه داده می‌گوییم، با استفاده از یک الگوریتم خوشه‌بندی مانند K-means یا DBSCAN [10] خوشه‌بندی کرده و پس از خلاصه‌سازی، داده‌ها را حذف می‌کنیم. هر خلاصه خوشه شامل مرکز، شعاع و تعداد نقاط داده مربوط به هر یک از رده‌ها است. شعاع یک خوشه، فاصله بین مرکز و دورترین نقطه داده درون خوشه در نظر گرفته می‌شود.

پس از حذف داده‌ها با ورود هر داده از جریان داده ورودی و با توجه به ویژگی‌های داده و شعاع خوشه، بررسی می‌شود که داده درون خوشه قرار می‌گیرد یا خیر. در صورتی که داده درون یکی از خوشه‌ها قرار گیرد، رأی بیشینه روی خوشه گرفته شده و رده با بیشترین تعداد تکرار در خوشه یادشده، به عنوان رده داده جریان در آن مدل تلقی می‌شود. حال با توجه به گروهی بودن روش رده‌بندی، دوباره بر روی مدل‌هایی که داده را درون خوشه تشخیص داده‌اند، رأی بیشینه گرفته شده و رده با بیشترین تکرار در مدل‌های موجود، به عنوان رده داده جریان در نظر گرفته می‌شود. در صورتی که داده برای هیچ یک از مدل‌ها درون خوشه‌ای قرار نگیرد، با استفاده از روشی مبتنی بر آنتروپی تشخیص تغییر برای نمونه بررسی شده و در صورتی که تغییر تشخیص داده شد، مدل گروهی به روزرسانی می‌شود.

در این مقاله از الگوریتم خوشه‌بندی K-means برای پیاده‌سازی استفاده شده و مقدار k بر اساس حجم داده‌های آموزشی تعیین می‌شود. داده‌های آموزشی به تعداد مدل‌هایی که در رده‌بند گروهی موجود است، تقسیم می‌شود. برای پیاده‌سازی این الگوریتم، داده‌های آموزشی

به شش قطعه داده تقسیم شده و روی هر قطعه i با استفاده از الگوریتم K-means خوشه‌بندی انجام می‌شود و مدل m_i به دست می‌آید. مقدار k برای هر ۶ قطعه داده یکسان در نظر گرفته شده است. پس از حذف داده‌ها، هر مدل m_i شامل مجموعه‌ای از k خلاصه داده است.

یک داده جریان x با استفاده از مدل m_i به صورت زیر رده‌بندی می‌شود. فرض کنید $h \in m_i$ یک خلاصه داده باشد که مرکز آن، نزدیک‌ترین مرکز به نمونه x نسبت به دیگر خلاصه‌های موجود در m_i است. با فرض این که مقدار k برای اجرای الگوریتم K-NN یک در نظر گرفته شود، رده‌ای که برای x در نظر گرفته می‌شود، بیشترین تعداد تکرار رده در خلاصه داده h است. رده‌بندی داده x با استفاده از رده‌بند گروهی M و با گرفتن رأی بیشینه از بین همه رده‌بندها انجام می‌شود. در صورتی که مقدار k غیر یک باشد، باید از بین k خلاصه داده رأی بیشینه گرفته شود. مرز تصمیم برای هر مدل m_i اجتماعی از فضای ویژگی همه خلاصه داده‌های $h \in m_i$ است. به همین ترتیب مرز تصمیم رده‌بند گروهی M به صورت اجتماعی از مرزهای تصمیم همه مدل‌های $m_i \in M$ تعریف می‌شود.

۳-۲- تشخیص تغییر

یک جریان داده توالی از چندتایی‌هایی است که هر نمونه داده x_i شامل بردار ویژگی f_i و یک برچسب z_i است و به شکل کلی (f_i, z_i) نمایش داده می‌شود. در این نمایش f_i برداری از همه ویژگی‌های نمونه z_i ام بوده و z_i برچسب نمونه z_i ام است. دو پنجره B_1 و B_2 با طول ثابت N در نظر می‌گیریم. با گذر جریان داده، تعداد N داده ابتدایی به صورت متوالی وارد پنجره اول و N داده بعدی وارد پنجره دوم می‌شود. با پرشدن پنجره‌ها، داده‌های بعدی یک‌به‌یک به پنجره B_2 وارد شده و از طرف مقابل قدیمی‌ترین داده از پنجره B_1 خارج می‌شود.

آنتروپی در نظریه اطلاعات معیاری عددی از میزان تصادفی بودن یک متغیر تصادفی است که به صورت رابطه (۱) تعریف شده است [15]. آنتروپی به دلیل ویژگی‌های مطلوب موجود و به ویژه وجود تقارن در آن، به عنوان یک معیار مناسب برای بررسی محتوای اطلاعات به کار گرفته می‌شود. برای مقایسه توزیع داده‌ها، آنتروپی نمونه‌های پنجره B_1 و پنجره B_2 را با هم مقایسه می‌کنیم. اگر اختلاف آنتروپی دو پنجره از حد مشخص λ کمتر بود، جریان داده پایدار بوده و هیچ تغییری رخ نداده است؛ ولی

اگر اختلاف توزیع داده‌ها در دو پنجره از λ بیشتر بود، تغییر تشخیص داده می‌شود. برای محاسبه آنتروپی جریان داده از رابطه (۲) استفاده می‌کنیم:

$$H(x) = -\sum_{i=1}^n P_i \log_2 P_i \quad (۱)$$

$$H(x) = -\sum_{i=1}^c \sum_x P_i^x \log_2 P_i^x \quad (۲)$$

که در این روابط x متغیر تصادفی گسسته، P_i احتمال رخداد α_i تابع احتمال چگالی x و c تعداد رده‌ها است. برای محاسبه P_i^x باید در هر مدل، خوشه‌ای که داده جریان x در آن واقع شده است و همچنین با توجه به تعداد نمونه کلاس i نسبت به کل نمونه‌های موجود در آن خوشه محاسبه شود؛ سپس از بین مقادیر به‌دست‌آمده برای هر مدل، میانگین گرفته و مقدار P_i^x محاسبه می‌شود.

۳-۳- به‌روزرسانی مدل گروهی رده‌بند

زمانی که یک تغییر مفهوم شناسایی می‌شود، رده‌بند با استفاده از داده‌های دو پنجره به‌روز می‌شود. به این صورت که یک مدل جدید با استفاده از داده‌های دو پنجره ساخته شده و جایگزین قدیمی‌ترین مدل از میان مدل‌های موجود می‌شود. این کار تضمین می‌کند به‌طور دقیق تعداد مشخصی مدل در هر زمان وجود داشته باشد. پس از به‌روزرسانی مدل، داده‌های دو پنجره حذف می‌شود. به این ترتیب روش پیشنهادی از داده‌های جریان نیز برای آموزش رده‌بند استفاده می‌کند و مسأله تغییر مفهوم هم با به‌روزرسانی مدل گروهی با استفاده از مفاهیم اخیر پاسخ داده می‌شود. زمانی که توزیع داده‌های یک رده تغییر کند، تغییر مفهوم رخ می‌دهد؛ بنابراین لازم است با به‌روزرسانی رده‌بند داده‌هایی که در قبل به‌عنوان داده پرت شناخته می‌شده، اکنون به‌عنوان نمونه‌ای از یک رده تلقی کنیم. این شناسایی تغییرات در بهبود عملکرد رده‌بندی مؤثر است. در صورتی که تغییری در توزیع داده‌ها اتفاق نیافتد، رده‌بند قادر است با مدل‌های موجود رده‌بندی را انجام دهد. به این دلیل به‌روزرسانی رده‌بند تنها زمانی که یک تغییر مفهوم شناسایی شود، انجام می‌شود.

۳-۴- پیچیدگی محاسباتی

در این بخش به تحلیل زمان اجرای الگوریتم پیشنهادی می‌پردازیم. در مرحله ساخت مدل اولیه، بر روی داده‌های آموزشی برچسب‌دار، الگوریتم خوشه‌بندی K-means اجرا می‌شود که زمان اجرای آن $O(tkn_l)$ است که n_l تعداد

داده‌های آموزشی اولیه است و t تعداد تکرار الگوریتم است. باید توجه کرد که مجموعه داده اولیه به چند قسمت تقسیم شده و الگوریتم خوشه‌بندی روی هر قسمت اجرا می‌شود. اما به‌صورت تجمیعی مرتبه زمان اجرا تغییری نمی‌کند و t برابر حداکثر تکرار الگوریتم در قسمت‌های مختلف مجموعه داده خواهد بود.

در هنگام ورود هر داده جدید، لازم است که ابتدا با مدل رده‌بند گروهی مورد ارزیابی قرار گیرد. بنابراین فاصله هر داده با خلاصه‌های نگهداری شده در هر مدل m_i مقایسه خواهد شد که تعداد این خلاصه‌ها حداکثر برابر k است. با فرض ثابت بودن تعداد مدل‌ها، این مقایسه برای هر داده از مرتبه $O(k)$ خواهد بود. محاسبات آنتروپی در پنجره‌های B_1 و B_2 به‌صورت افزایشی با مرتبه ثابت انجام می‌شود. در صورت تشخیص تغییر یک مدل جدید با اجرای خوشه‌بندی K-means بر روی داده‌های دو پنجره با مرتبه زمانی $O(tk(|B_1| + |B_2|))$ ایجاد می‌شود که t تعداد تکرار در خوشه‌بندی است. بنابراین در بدترین حالت هر داده جریان در یک اجرای الگوریتم K-means دخیل بوده و به‌ازای هر داده از مرتبه $O(tk)$ کار انجام می‌شود.

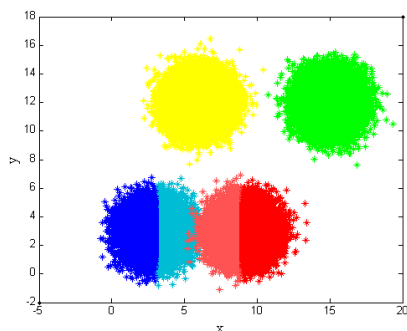
۴- پیاده‌سازی و نتایج

در این بخش، به بیان مجموعه داده و مقایسه روش پیشنهادی با برخی از روش‌های موجود در رده‌بندی جریان داده پرداخته می‌شود. پیاده‌سازی این مقاله به زبان جاوا انجام شده است.

۴-۱- مجموعه داده

بخشی از مجموعه داده‌های استفاده‌شده برای این مقاله از تارنمای UCI به‌دست آمده است. در جدول (۱) مشخصات مجموعه داده‌های استفاده‌شده برای ارزیابی روش پیشنهادی با سایر روش‌های رده‌بندی آورده شده است. مجموعه داده Gaussian Generator با استفاده از توزیع گوسی با چهار رده و دو ویژگی ساخته شده که مطابق شکل (۳) تعدادی از نمونه‌های دو رده آبی و قرمز هم‌پوشانی دارند. برای ایجاد تغییر مفهوم، داده‌های سمت چپ رده آبی و سمت راست رده قرمز به‌عنوان داده‌های آموزشی و برای داده‌های جریان، داده‌های سمت راست رده آبی و سمت چپ رده قرمز به الگوریتم داده شده است. برای دو رده سبز و زرد رنگ هم بخشی از نمونه‌های این دو رده به‌صورت تصادفی به‌عنوان داده آموزشی در نظر گرفته شده و بقیه داده‌های آن‌ها به‌عنوان داده جریان به

مدل‌های اولیه در نظر گرفته شده و بقیه نمونه‌ها به‌عنوان داده‌های جریان به الگوریتم داده شده است.



(شکل-۳): نمایش مجموعه داده ساختگی با استفاده از

توزیع گاوسی

(Figur-3): The Gaussian generator artificial data set

۴-۲- معیار ارزیابی

برای نشان دادن عملکرد روش پیشنهادی در رده‌بندی جریان داده از معیارهای صحت^۲ (Accu)، دقت^۳ (PR)، یادآوری^۴ (RE) و F1 استفاده شده که در زیر معرفی شده‌اند.

$$Accu(y, \hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y}_i = y_i) \quad (۳)$$

$$PR = \frac{TP}{TP+FP} \quad (۴)$$

$$RE = \frac{TP}{TP+FN} \quad (۵)$$

$$F1 = \frac{2*PR*RE}{PR+RE} \quad (۶)$$

(جدول-۲): مقایسه روش پیشنهادی و دو روش OBA و HAT

بر روی مجموعه داده Forest Cover

(Table-2): Comparison of the proposed method OBA and HAT on the Forest Cover data set

HAT	OBA	روش پیشنهادی	معیار / روش
59.8	66.4	99.3	C1 PR
66	68.99	66.04	C2 PR
77.98	79.65	100	C3 PR
47.33	48.36	100	C4 PR
17.53	17.76	100	C5 PR
28.94	30.59	100	C6 PR
55.66	53.98	100	C7 PR
57.09	45.91	53.52	C1 RE
79.2	89.94	100	C2 RE
52.07	56.69	35.19	C3 RE
39.96	42.21	5.46	C4 RE
14.56	9.05	52.15	C5 RE
74.8	80.64	36.23	C6 RE
59.06	47.53	60.61	C7 RE
58.41	54.29	69.55	C1 F1
72	78.08	79.55	C2 F1
62.44	66.24	52.06	C3 F1
43.33	45.08	10.35	C4 F1
15.91	11.99	68.55	C5 F1
41.73	44.35	53.19	C6 F1
57.31	50.55	75.47	C7 F1

² Accuracy

³ Precision

⁴ Recall

الگوریتم داده شده است. هدف استفاده از این مجموعه داده، بررسی عملکرد الگوریتم با وجود تغییر مفهوم و هم‌پوشانی دو رده به صورت هم‌زمان است. مجموعه داده Forest Cover شامل ۵۴ ویژگی و هفت رده بوده که ما برای ارزیابی نتایج، پس از نرمال‌سازی از صدهزار نمونه آن استفاده کرده‌ایم.

(جدول-۱): مشخصات مجموعه داده‌های استفاده شده برای

ارزیابی عملکرد روش پیشنهادی

(Table-1): The characteristics of data sets used to evaluate performance of the proposed method

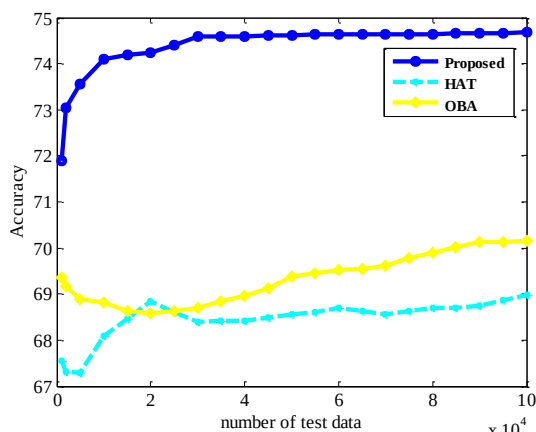
نام مجموعه داده	تعداد نمونه	تعداد کلاس	تعداد ویژگی
Forest Cover	100,000	7	54
Shuttle	40,000	7	9
Skin Segmentation	100,000	2	3
Gaussian Generator	150,000	4	2
Random RBF Generator Drift	100,000	5	17

مجموعه داده Shuttle شامل ۷ ویژگی و هفت رده است که چهل هزار نمونه برای ارزیابی نتایج انتخاب شده است. در مجموعه داده Skin Segmentation از صدهزار نمونه به‌عنوان داده آموزشی و داده جریانی استفاده شده است. این سه مجموعه داده از تارنمای UCI گرفته شده و مجموعه داده آخر با استفاده از Random RBF Generator در Drift با پنج رده و هفده ویژگی ساخته شده است. MOA یک محیط نرم‌افزاری برای اجرای الگوریتم‌ها و اجرای آزمایش‌ها برای یادگیری برخط^۱ جریان داده است [16]. مجموعه داده Random RBF Generator به این صورت ساخته می‌شود که ابتدا تعدادی مرکز تصادفی تولید می‌شود. به طوری که هر مرکز دارای موقعیت، انحراف معیار، برچسب رده و وزن است. نمونه‌های جدید با انتخاب تصادفی یک مرکز تولید شده و وزن‌ها برای احتمال انتخاب مراکز، اختصاص می‌یابند. مرکز انتخاب‌شده همچنین برچسب رده نمونه را تعیین می‌کند. تغییر مفهوم با تغییر در مراکز و با یک سرعت ثابت در مجموعه داده ایجاد می‌شود [14]. در مجموعه داده Random RBF Generator Drift از MOA برای ساخت داده‌ها استفاده شده و سرعت تغییر مراکز ۰/۰۰۱ واحد در نظر گرفته شده است.

در تمامی مجموعه داده‌های معرفی شده، تعداد شش هزار نمونه به‌عنوان داده آموزشی برای ساخت

¹ Online

مشخص است، روش پیشنهادی در تمام مراحل به مراتب از دو روش دیگر صحت بیشتری داشته و عملکرد خوبی را ارائه داده است.



(شکل-۴): مقادیر تغییر صحت به ازای اضافه شدن پنج هزار

داده جریان در هر مرحله برای مجموعه داده Forest Cover
(Figure-4): Accuracy values while adding 5,000 test data per step for the Forest Cover data set

در شکل (۵) روش پیشنهادی با دو روش OBA و HAT بر روی مجموعه داده Shuttle مقایسه شده که در هر مرحله دوهزار داده به مجموعه داده جریان اضافه شده است. همان طور که مشاهده می شود، صحت روش پیشنهادی در ابتدا با تعداد داده جریان کمتر از دوهزار با روش OBA برابر است؛ ولی با زیاد شدن داده های جریان خیلی سریع صحت روش پیشنهادی بالا رفته و از روش OBA فاصله می گیرد. بالا رفتن صحت روش پیشنهادی برای هر دو مجموعه داده، نشان دهنده آن است که بیشتر تغییرات موجود در این دو مجموعه داده از نوع تدریجی و ناگهانی بوده و این روش به خوبی تغییرات موجود در داده ها را تشخیص داده و از آن ها برای آموزش الگوریتم به روزرسانی مدل گروهی رده بند استفاده می کند.

در جدول (۳) صحت روش پیشنهادی در مقایسه با دو روش OBA و HAT و همچنین روش پیشنهادی مرجع [23] بر روی مجموعه داده های مختلف نشان داده شده است. مرجع [23] از دو ماتریس خلاصه برای تقریب داده ها استفاده می کند و برای هر ماتریس با استفاده از تجزیه مقادیر منفرد، بهینه ترین تقریب را به دست می آورد. همچنین از میانگین اعداد به دست آمده از این ماتریس به عنوان معیار شباهت استفاده کرده و برای تشخیص رده جدید بهره می برد. پیچیدگی زمانی روش پیشنهادی و روش [23] هر دو از مرتبه خطی نسبت به تعداد داده ها است. پیچیدگی زمانی دو روش دیگر در مراجع یاد نشده است. همان طور که مشاهده می شود، در تمامی مجموعه داده ها روش پیشنهادی از صحت بالاتری نسبت به

در رابطه (۳) منظور از $\hat{y}_i$ مقدار پیش بینی شده نمونه i ام و y_i مقدار واقعی نمونه i و n تعداد نمونه های موجود است. در رابطه (۴) و (۵) منظور از TP تعداد نمونه هایی است که به درستی به وسیله رده بند به رده C_i انتساب یافته و FP یعنی تعداد نمونه هایی که متعلق به سایر رده ها بوده و به اشتباه به وسیله رده بند به رده C_i منسوب شده اند. FN تعداد نمونه هایی است که متعلق به رده C_i بوده ولی به وسیله رده بند به سایر رده ها تعلق گرفته است. معیار TN تعداد نمونه هایی است که متعلق به کلاس C_i نبوده اند و به وسیله رده بند به این رده اختصاص داده نشده اند. در رابطه (۶) هم معیار F1 با استفاده از دو معیار دقت و یادآوری به دست می آید.

۳-۴- نتایج آزمایش ها

ابتدا مقایسه ای بین دو روش OBA و HAT با روش پیشنهادی انجام شده است. روش HT (Hoeffding Tree) همان الگوریتم درخت تصمیم است که با اعمال تغییراتی برای اجرا بر روی جریان داده سازگار شده است. ایده اصلی این الگوریتم محدود کردن سطح اطمینان بهترین ویژگی برای تقسیم درخت است. روش HAT (Hoeffding Adaptive Tree) با ایجاد تغییر در بیشینه تعداد تقسیم و حذف برخی گره های درخت در روش HT، سعی در کوچکتر کردن درخت برای سازگاری سریع تر الگوریتم با تغییرات موجود در جریان داده کرده است [17]. OBA (OzaBagAdwin) یک روش گروهی برای تشخیص تغییر به همراه تخمین وزن برای بهبود عملکرد است. تشخیص تغییر با مقایسه میانگین مقادیر نمونه هایی که قبل و بعد از اضافه شدن نمونه i ام به پنجره به دست می آیند، انجام می گیرد. زمانی که تغییر، تشخیص داده می شود، رده بند جدید جایگزین بدترین رده بند در مدل گروهی می شود [14].

جدول (۲) نتایج به دست آمده از رده بندی روش پیشنهادی و دو روش OBA و HAT را بر روی مجموعه داده Forest Cover نشان می دهد. در این جدول میزان دقت، یادآوری و همچنین F1 هر کدام از این سه روش به تفکیک برای هر رده آمده است. همان طور که مشاهده می شود، روش پیشنهادی برای بیشتر رده های موجود در مجموعه داده در هر سه معیار ارزیابی بالاتر بوده و در کل عملکرد بهتری داشته است.

در شکل (۴) مقایسه این سه روش بر روی مجموعه داده Forest Cover با معیار ارزیابی صحت نشان داده شده که در هر مرحله پنج هزار داده به مجموعه داده جریان اضافه شده است. همان طور که از شکل (۴)

مجموعه داده‌ها است. به عنوان کارهای آینده می‌توان تشخیص مفاهیم جدید را به روش ارائه شده اضافه کرد. بخشی از داده‌های پرتی که به وسیله الگوریتم ارائه شده کنار گذاشته می‌شوند، در واقع داده‌هایی هستند که نمایانگر یک مفهوم جدید هستند. با ارائه روشی می‌توان این مفاهیم جدید را تشخیص داده و دقت رده‌بند را افزایش داد.

سپاس و قدردانی

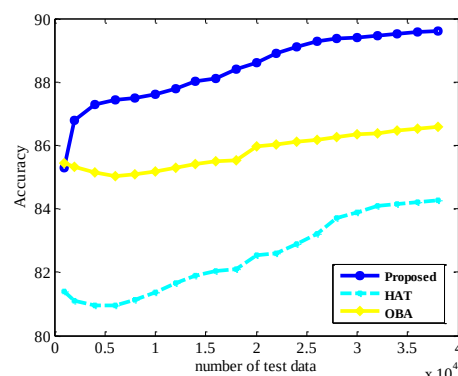
به این وسیله از بنیاد ملی نخبگان جهت حمایت از پژوهش سپاس‌گزاری و قدردانی به عمل می‌آید.

6- References

۶- مراجع

- [1] M. Masud, J. Gao, L. Khan, J. Han and B. M. Thuraisingham, "Classification and novel class detection in concept-drifting data streams under time constraints," *IEEE Transactions on knowledge and data engineering*, vol. 23, no. 6, pp. 859-874, 2010.
- [2] M. M. Masud, Q. Chen, L. Khan, C. Aggarwal, J. Gao, J. Han and B. Thuraisingham, "Addressing concept-evolution in concept-drifting data streams," in *2010 IEEE International Conference on Data Mining*, IEEE, 2010, pp. 929-934.
- [3] B. S. Parker and L. Khan, "Detecting and tracking concept class drift and emergence in non-stationary fast data streams," in *Twenty-ninth AAAI conference on artificial intelligence*, 2015.
- [4] R. Klinkenberg, "Learning drifting concepts: Example selection vs. example weighting," *Intelligent data analysis*, vol. 8, no. 3, pp. 281-300, 2004.
- [5] A. Bifet and R. Gavalda, "Learning from time-changing data with adaptive windowing," in *Proceedings of the 2007 SIAM international conference on data mining*, SIAM, 2007, pp. 443-448.
- [6] A. Haque, L. Khan and M. Baron, "Sand: Semi-supervised adaptive novel class detection and classification over data stream," in *THIRTIETH AAAI Conference on Artificial Intelligence*, 2016.
- [7] L. I. Kuncheva and W. J. Faithfull, "PCA feature extraction for change detection in multidimensional unlabeled data," *IEEE transactions on neural networks and learning systems*, vol. 25, no. 1, pp. 69-80, 2013.
- [8] P. Sidhu and M. Bhatia, "A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority," *International Journal of Machine Learning and Cybernetics*, vol. 9, no. 1, pp. 37-61, 2018.

دو روش دیگر برخوردار بوده و عملکرد بهتری دارد. این نشان می‌دهد که بیش‌تر تغییرات موجود در داده‌ها از نوع تدریجی و ناگهانی است و این روش به خوبی توانسته تغییرات موجود در مجموعه داده را تشخیص داده و با به‌روزرسانی مدل گروهی رده‌بند، در بهبود دقت و صحت الگوریتم مؤثر باشد.



(شکل ۵): مقادیر تغییر صحت به‌ازای اضافه‌شدن دوهزار داده

جریان در هر مرحله برای مجموعه داده Shuttle

(Figure-5): Accuracy values while adding 2,000 test data per step for the Shuttle data set

(جدول ۳): مقایسه صحت روش‌ها بر روی

مجموعه داده‌های مختلف

(Table-3): Comparing accuracy of the methods on different data sets

نام مجموعه داده	روش پیشنهادی	OBA	HAT	[23]
Forest Cover	74.67	70.16	68.98	71.05
Shuttle	89.61	86.57	84.26	87.79
Skin Segmentation	100	99.4	99.21	100
Gaussian Generator	98.3	95.2	95.1	97.04
Random RBF Generator Drift	81.32	79.83	78.06	80.07

۵- نتیجه‌گیری

در این مقاله، یک رده‌بند گروهی نیمه‌نظارتی معرفی شده که می‌تواند دو نوع تغییر مفهوم تدریجی و ناگهانی را در جریان داده تشخیص داده و دقت رده‌بندی را بهبود بخشد. این رده‌بند گروهی برای تشخیص تغییر از آنتروپی استفاده می‌کند، به‌طوری‌که با وارد شدن هر داده تغییر آنتروپی داده‌ها بررسی و در صورتی که تغییر از حد معینی بیشتر بود، تغییر مفهوم تشخیص داده می‌شود. با تشخیص تغییر مفهوم یک مدل جدید از داده‌های اخیر ساخته شده و مدل گروهی به‌روزرسانی می‌شود. این رده‌بند بر روی مجموعه داده‌های مختلفی با سه روش موجود مقایسه شده است. نتایج به‌دست آمده نشان‌دهنده عملکرد بهتر روش پیشنهادی نسبت به روش‌های دیگر بر روی این

- Fuzzy Systems, vol. 28, no. 3, pp. 544-557, 2019.
- [22] Y. Li, Y. Wang, Q. Liu, C. Bi, X. Jiang, and S. Sun. "Incremental semi-supervised learning on streaming data." *Pattern Recognition*, vol. 88 pp. 383-396, 2019.
- [23] X. Mu, F. Zhu, J. Du, E.P. Lim, & Z.H. Zhou, "Streaming classification with emerging new class by class matrix sketching" In Thirty-First AAAI Conference on Artificial Intelligence, pp. 2373-2379, 2017.
- [24] P. Vorburger, A. Bernstein. "Entropy-based concept shift detection" In Sixth IEEE International Conference on Data Mining, ICDM'06, pp. 1113-1118, 2006.
- [25] L. Du, Q. Song, and X. Jia. "Detecting concept drift: an information entropy based method using an adaptive sliding window." *Intelligent Data Analysis* vol. 18, no. 3, pp. 337-364, 2014.
- [26] J. Haug, G. Kasneci. "Learning Parameter Distributions to Detect Concept Drift in Data Streams". arXiv preprint arXiv:2010.09388. 2020.
- [27] H. Hanqing, M. Kantardzic, T. S. Sethi. "No Free Lunch Theorem for concept drift detection in streaming data classification: A review." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol. 10, no. 2, e1327, 2020.
- [۲۸] م. مسافری، ع. صفائی. ارائه روشی پویا جهت پاسخ به پرس‌وجوهای پیوسته تجمعی اقتضایی. پردازش علائم و داده‌ها. جلد ۱۴، شماره ۳، ص ۲۲-۳، ۱۳۹۶.
- [23] M. Mosaferi, A. Safaei, "Providing a Dynamic Technique for Answering Ad-hoc Continuous Aggregate". *Journal of Signal and Data Processing*. Vol. 14, No. 3, pp. 3-22, 2017.
- حسین حسن‌نژاد نامقی مدرک کارشناسی خود را در رشته کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی شاهرود در سال ۱۳۹۴ دریافت کرد. وی در حال حاضر دانشجوی مقطع کارشناسی ارشد در دانشگاه صنعتی شاهرود در گرایش هوش مصنوعی است. موضوع پایان‌نامه کارشناسی ارشد ایشان، "رده‌بندی نیمه‌نظارتی جریان‌های داده تکاملی" است. زمینه‌های پژوهشی مورد علاقه ایشان داده‌کاوی، یادگیری ماشین و کاوش داده‌های حجیم است. نشانی رایانامه ایشان عبارت است از: h_hasannezhad@shahroodut.ac.ir
- [9] O. A. Mahdi, E. Pardede and J. Cao, "Combination of information entropy and ensemble classification for detecting concept drift in data stream," in *Proceedings of the Australasian Computer Science Week Multiconference*, ACM, 2018, p. 13.
- [10] M. Ester, H.-P. Kriegel, J. Sander and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Kdd*, 1996, pp. 226-231.
- [11] X. Zhu and A. B. Goldberg, "Introduction to semi-supervised learning," *Synthesis lectures on artificial intelligence and machine learning*, vol. 3, no. 1, pp. 1-130, 2009.
- [12] A. Tsymbal, "The problem of concept drift: definitions and related work," *Computer Science Department, Trinity College Dublin*, vol. 106, no. 2, p. 58, 2004.
- [13] I. Žliobaitė, "Learning under concept drift: an overview," in *arXiv preprint arXiv:1010.4784*, 2010.
- [14] A. Bifet, G. Holmes, B. Pfahringer, R. Kirkby and R. Gavalda, "New ensemble methods for evolving data streams," in *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, ACM, 2009, pp. 139-148.
- [15] S. J. Morshed, J. Rana and M. Milrad, "Real-time Data analytics: An algorithmic perspective," in *International Conference on Data Mining and Big Data*, Springer, 2016, pp. 311-320.
- [16] A. Bifet, G. Holmes, R. Kirkby and B. Pfahringer, "Moa: Massive online analysis," *Journal of Machine Learning Research*, vol. 11, no. May, pp. 1601-1604, 2010.
- [17] B. Pfahringer, G. Holmes and R. Kirkby, "Handling numeric attributes in hoeffding trees," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Berlin, Heidelberg, Springer, 2008, pp. 296-307.
- [18] D.L. Cabral, D. Rafael, and R.S.M. de Barros. "Concept drift detection based on Fisher's Exact test." *Information Sciences*, vol. 442, pp. 220-234, 2018.
- [19] R.F. de Mello, Y. Vaz, C.H. Grossi, and A. Bifet. "On learning guarantees to unsupervised concept drift detection on data streams." *Expert Systems with Applications*. Vol. 117, pp. 90-102, 2019.
- [20] X. Wang, Q. Kang, M. Zhou, L. Pan, and A. Abusorrah. "Multiscale Drift Detection Test to Enable Fast Learning in Nonstationary Environments." *IEEE Transactions on Cybernetics*, pp. 1-13, 2020.
- [21] Y. Song, J. Lu, H. Lu, and G. Zhang. "Fuzzy clustering-based adaptive regression for drifting data streams." *IEEE Transactions on*



حسین حسن‌نژاد نامقی مدرک کارشناسی خود را در رشته کامپیوتر گرایش نرم‌افزار از دانشگاه صنعتی شاهرود در سال ۱۳۹۴ دریافت کرد. وی در حال حاضر دانشجوی مقطع کارشناسی ارشد در دانشگاه صنعتی شاهرود در گرایش هوش مصنوعی است. موضوع پایان‌نامه کارشناسی ارشد ایشان، "رده‌بندی نیمه‌نظارتی جریان‌های داده تکاملی" است. زمینه‌های پژوهشی مورد علاقه ایشان داده‌کاوی، یادگیری ماشین و کاوش داده‌های حجیم است. نشانی رایانامه ایشان عبارت است از:

h_hasannezhad@shahroodut.ac.ir

هدی مشایخی فارغ التحصیل مقطع



دکترای تخصصی رشته مهندسی کامپیوتر، گرایش نرم افزار از دانشگاه صنعتی شریف است. پیش از آن، مقاطع کارشناسی و کارشناسی ارشد را

نیز در همان دانشگاه به پایان رسانیده است. وی در حال حاضر استادیار دانشگاه صنعتی شاهرود است. داده کاوی و یادگیری در زمینه داده های حجیم و پردازش توزیع شده از جمله علایق پژوهشی ایشان است.

نشانی رایانامه ایشان عبارت است از:

hmashayekhi@shahroodut.ac.ir

مرتضی زاهدی در حال حاضر عضو



هیأت علمی دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود است. وی پروژه هایی را در زمینه تعامل انسان و رایانه، شناسایی الگو، پردازش تصویر و

ویدئو، و بینایی ماشین در دست اجرا دارد که در آنها به طور معمول از اطلاعات و دانش آماری استفاده می شود. ایشان دارای دکترای تخصصی کامپیوتر از دانشگاه RWTH-Aachen آلمان است. تألیف کتب و مقالات علمی و همچنین سرپرستی پروژه های دانشگاهی و صنعتی در ایران و کشورهای اروپایی در کارنامه کاری ایشان دیده می شود.

نشانی رایانامه ایشان عبارت است از:

zahedi@suigle.com