
بسم الله الرحمن الرحيم

دانشکده مهندسی کامپیوتر

پایان‌نامه کارشناسی ارشد مهندسی هوش مصنوعی

رده‌بندی نیمه‌نظارتی جریان‌های داده تکاملی

نگارنده: حسین حسن نژاد نامقی

استاد راهنما

دکتر هدی مشایخی

استاد مشاور

دکتر مرتضی زاهدی

تیر ۱۳۹۸

شماره: ۵۷۹ /ف.ک
تاریخ: ۹۸/۴/۲۲

باسمه‌تعالی

دانشگاه صنعتی شاهرود

مدیریت تحصیلات تکمیلی

فرم شماره (۳) صورتجلسه نهایی دفاع از پایان نامه دوره کارشناسی ارشد

با نام و یاد خداوند متعال، ارزیابی جلسه دفاع از پایان نامه کارشناسی ارشد آقای حسین حسن نژاد نامقی با شماره دانشجویی **۹۵۰۴۵۴۴** رشته مهندسی کامپیوتر گرایش هوش مصنوعی تحت عنوان رده‌بندی نیمه‌نظارتی جریان‌های داده تکاملی که در تاریخ ۱۳۹۸/۴/۲۲ با حضور هیأت محترم داوران در دانشگاه صنعتی شاهرود برگزار گردید به شرح ذیل اعلام می‌گردد:

تقدیم اول به

پدر و مادر عزیزم که مهر آسمانی شان آرامش بخش

و راه و روش شان چراغ راه زندگانی ام است

و تقدیم دوم

به همه افرادی ست که درست اندیشیدن را در این مسیر به من آموختند.

تشکر و قدردانی

سپاس خداوند یکتای عزتمندی که رحمت و دانش او در سراسر گیتی گسترده شده، آسمان ها و زمین همه از آن اوست و علم و دانش حقیقی را بر هر که بخواهد موهبت می فرماید. رحمت و لطف او را بی نهایت سپاس می گویم چرا که فهم و درک مطالب این پژوهش را بر من ارزانی داشت و مرا به این اصل رساند که علم و ایمان دو بال یک پرواز ند. توفیق تلاش به من داد، تا با امید، راه تازه ای را آغاز کنم و به خواست او به نتیجه ی مطلوب نائل آیم. به راستی که همه چیز از آن و به خواست اوست. همچنین از استاد گرامی، سرکار خانم دکتر مهدی مشایخی بسیار سپاسگزارم که در تمامی دشواری های این مسیر، راهنمایی های بی دریغشان چاره ساز کارم بود و از آقای دکتر مرتضی زاهدی بابت کمک هایشان قدردانی به عمل می آورم.

تعهدنامه

اینجانب **حسین حسن نژاد نامقی** دانشجوی دوره کارشناسی ارشد رشته **مهندسی کامپیوتر** دانشکده مهندسی کامپیوتر دانشگاه صنعتی شاهرود نویسنده پایان‌نامه **رده‌بندی نیمه‌نظارتی جریان‌های داده تکاملی** تحت راهنمائی **دکتر هدی مشایخی** متعهد می‌شوم.

- تحقیقات در این پایان‌نامه توسط اینجانب انجام شده است و از صحت و اصالت برخوردار است .
- در استفاده از نتایج پژوهشهای محققان دیگر به مرجع مورد استفاده استناد شده است .
- مطالب مندرج در پایان‌نامه تاکنون توسط خود یا فرد دیگری برای دریافت هیچ نوع مدرک یا امتیازی در هیچ جا ارائه نشده است .
- کلیه حقوق معنوی این اثر متعلق به دانشگاه صنعتی شاهرود می‌باشد و مقالات مستخرج با نام « دانشگاه صنعتی شاهرود » و یا « Shahrood University of Technology » به چاپ خواهد رسید .
- حقوق معنوی تمام افرادی که در به دست آمدن نتایج اصلی پایان‌نامه تأثیرگذار بوده اند در مقالات مستخرج از پایان‌نامه رعایت می گردد.
- در کلیه مراحل انجام این پایان‌نامه ، در مواردی که از موجود زنده (یا بافتهای آنها) استفاده شده است ضوابط و اصول اخلاقی رعایت شده است .
- در کلیه مراحل انجام این پایان‌نامه، در مواردی که به حوزه اطلاعات شخصی افراد دسترسی یافته یا استفاده شده است اصل رازداری ، ضوابط و اصول اخلاق انسانی رعایت شده است .

تاریخ

امضای دانشجو

مالکیت نتایج و حق نشر

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، عموما فرض می‌شود تنها بخشی از داده‌ها دارای برچسب هستند. همچنین یکی دیگر از چالش‌های مهم تحلیل جریان داده تکامل مفهوم است. زمانی که تغییر داده‌ها موجب پدیدار شدن مفاهیم جدید می‌گردد، تکامل مفهوم اتفاق می‌افتد و به صورت ظاهر شدن کلاس جدید نمایان می‌شود. 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 به روز رسانی مدل رده‌بند گروهی استفاده می‌کند. همچنین در این روش کلاس جدید با شناسایی داده‌های دور افتاده‌ای که همبستگی قوی بین خودشان دارند، تشخیص داده می‌شود. روش پیشنهادی قادر است تغییرات موجود در مجموعه داده را تشخیص داده و با به روز رسانی مدل یادگیری، در بهبود دقت الگوریتم موثر باشد. نتایج آزمایش‌ها نشان می‌دهد که روش پیشنهادی از نظر تشخیص تغییر، تکامل مفهوم و دقت نسبت به سایر روش‌ها کارایی بیشتری دارد.

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

لیست مقالات مستخرج از پایان نامه

۱-حسن‌نژاد نامقی،ح. مشایخی،ه. (۱۳۹۸)، "تشخیص تغییر مفهوم در جریان داده با کمک رده‌بند نیمه‌نظارتی" پردازش علائم و داده‌ها (تحت داوری)

فهرست مطالب

فهرست جداول

فهرست اشکال

فصل۱ : مقدمه

۱-۱ مقدمه

پیشرفت‌های اخیر سخت‌افزاری در زمینه‌های ذخیره‌سازی و پردازش اطلاعات، امکان جمع‌آوری خودکار داده‌ها را فراهم کرده است. با افزایش حجم داده‌ها، به روش‌هایی جهت استخراج دانش از منابع بزرگ داده نیاز است. به این‌گونه 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. این جریان‌های داده در کاربردهای مختلفی همچون بررسی تراکنش‌های مالی، داده‌های بدست آمده از حسگرهای بی‌سیم، بررسی داده‌های ترافیک شبکه، داده‌های موجود در شبکه‌های اجتماعی و اطلاعات روزانه سبدهای خرید کارایی دارند [۱]. از چالش‌های موجود در این حوزه که موجب مجزا ساختن آن‌ها از داده‌های ایستا[۲] می‌گردد، می‌توان به اندازه نامحدود داده‌ها، پردازش برخط[۳]، وجود تغییر مفهوم، عدم دسترسی به کل داده‌ها و ماهیت افزایشی و پویا[۴] اشاره کرد [۲].

در کاوش جریان‌های داده از روش‌های گوناگونی همچون رده‌بندی[۵]، خوشه‌بندی[۶]، درخت تصمیم[۷] و شبکه‌های عصبی[۸] استفاده می‌گردد. ولی در این پژوهش تمرکز بیشتر بر روی رده‌بندی می‌باشد. تغییر مفهوم[۹] زمانی اتفاق می‌افتد که توزیع ویژگی یا برچسب داده‌ها در طول زمان تغییر کند [۱۲]. بنابراین مدل رده‌بند باید با توجه به توزیع جدید به صورت پیوسته به روز گردد. به طور کلی الگوریتم‌های رده‌بندی با در نظر گرفتن تغییر مفهوم به دو دسته الگوریتم‌های کور و آگاه تقسیم می‌شوند [۱۳]. در روش‌های کور بدون توجه به تغییر مفهوم، مدل در زمان‌هایی مشخص به روز می‌شود. ولی در روش‌های آگاه با بررسی جریان داده، در صورت رخداد تغییر مفهوم مدل به روز خواهد شد. همچنین روش‌های

---

[1] Data stream
[2] Static
[۳] Online
[4] Dynamic
[5] Classification
[6] Clustering
[7] Decision tree
[8] Neural networks
[9] Concept drift

یادگیری را می‌توان با توجه به میزان داده‌های آموزشی برچسب‌دار به سه دسته با نظارت، نیمه‌نظارتی و بدون نظارت دسته‌بندی 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 درصورتی که علاوه بر داده‌های برچسب‌دار، از داده‌های بدون برچسب در یادگیری استفاده شود، یادگیری نیمه‌نظارتی خواهد بود. برای جریان‌های داده، دسته نیمه‌نظارتی الگوریتم‌های یادگیری بیش از دسته نظارتی مورد نیاز است؛ چرا که همانطور که عنوان شد، در جریان‌های داده فرض وجود برچسب برای همه داده‌ها یک فرض غیر واقع‌بینانه بوده و تعداد داده‌های برچسب‌دار نسبت به کل داده‌ها اندک است [۱۱].

۱-۲ شرح مسئله

با وجود کارهای انجام گرفته در سال‌های اخیر، هنوز تحقیقات گسترده‌ای در زمینه رده‌بندی جریان‌های داده مورد نیاز است. اکثر الگوریتم‌های ارائه شده، با در نظر گرفتن بخشی از چالش‌ها سعی در برطرف کردن برخی از مشکلات داشته‌اند. امروزه مجموعه داده‌ها برای حافظه‌های اصلی بسیار بزرگ بوده و نیاز به ذخیره‌سازی در حافظه‌های جانبی دارند. بنابراین استفاده از روش‌های دسترسی تصادفی مطابق آنچه که در روش‌های سنتی داده کاوی بکار می‌رفت، بسیار هزینه‌بر است [۱۸].

استفاده از الگوریتم‌های متداول یادگیری به‌دلیل 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 [۱۱].

در پژوهش‌های مرتبط، برای مقابله با چالش نامحدود بودن طول جریان داده، از تقسیم جریان داده به قطعاتی با اندازه ثابت [۳]، یا فراموشی تدریجی[۱] [۴] استفاده شده است. برای مواجهه با چالش تغییر مفهوم از به روز رسانی دوره‌ای رده‌بند [۱]، یا تشخیص دهنده تغییر صریح [۵] برای تشخیص زمان به روز رسانی استفاده شده است. این رویکردها با فرض اینکه برچسب‌های درست برای تمام نمونه‌های داده موجود است، مطرح شده‌اند. درحالی‌که عموما 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 [۶]. بنابراین باید به دنبال راهی بود تا از داده‌های بدون برچسب نیز به عنوان داده‌های آموزشی در الگوریتم استفاده کرد.

در این پایان‌نامه روشی برای رده‌بندی تدریجی جریان داده ارائه شده که از آنتروپی به عنوان شناسایی کننده تغییر استفاده می‌کند. همچنین سعی شده است که با ارائه راهکاری مبتنی بر یادگیری نیمه‌نظارتی که به صورت افزایشی مدل خود را به روز رسانی می‌کند، بتواند پاسخگوی چالش‌های موجود در این حوزه از جمله اندازه نامحدود داده‌ها، پردازش برخط، وجود تغییر مفهوم، عدم دسترسی به کل داده‌ها و ماهیت افزایشی و پویا باشد. این روش یادگیری از رده‌بندهای گروهی استفاده می‌کند. رده‌بندهای گروهی می‌توانند به روز رسانی شده و تغییرات احتمالی جریان داده را مدیریت کنند. الگوریتم ارائه شده بر روی مجموعه داده‌های استاندارد آزمایش شده و با برخی از روش‌های موجود در این زمینه مقایسه می‌گردد. مقایسه روش پیشنهادی با در نظر گرفتن معیارهای مختلف کارایی از جمله صحت[۲]، دقت[۳]، یادآوری[۴] و F1 صورت خواهد گرفت. مقایسه‌های انجام شده، نشان دهنده کارایی بیشتر روش پیشنهادی نسبت به روش‌های موجود خواهد بود.

[1] Gradual forgetting
[۲] Accuracy
[۳] Percision
[۴] Recall

۱-۳ اهمیت انجام پژوهش

امروزه در بیشتر کاربردها و در محیط‌های پویا که اطلاعات در آن‌ها به مرور تولید می‌شوند، برای پردازش جریان‌های داده به الگوریتم‌های یادگیری نیاز دارند. با این وجود در سه دهه اخیر، تحقیقات انجام شده در زمینه داده‌کاوی بر روی یادگیری دسته‌ای[1] تمرکز داشته‌اند. در یادگیری دسته‌ای تمام داده‌های آموزشی در دسترس الگوریتم بوده و با میانگین‌گیری از مدل‌های حاصل از آموزش داده‌هایی که به صورت تصادفی برای این امر انتخاب شده‌اند، بر مشکل محدود بودن داده‌های آموزشی غلبه می‌شود [۴۱]. از سوی دیگر پویایی بسیاری از مسائل دنیای واقعی، تحقیقات در این زمینه از داده‌کاوی را ضروری ساخته است. در محیط‌های پویا داده‌ها در طول زمان جمع‌آوری شده و دیگر نمی‌توان از الگوریتم‌های یادگیری دسته‌ای استفاده کرد.

همانطور که عنوان شد، یکی از مهم‌ترین چالش‌های موجود در زمینه داده‌کاوی وجود تغییر مفهوم در آن‌هاست. به تدریج با زیاد شدن تغییر مفاهیم در جریان داده، دقت رده‌بندی کاهش می‌یابد. از این رو تشخیص این تغییرات و به روز رسانی مدل‌های رده‌بندی با توجه به این تغییرات امری ضروری است. روش‌های تشخیص تغییر مفهوم را می‌توان بر اساس تعداد رده‌بند، به روش‌های تک رده‌بند یا گروهی تقسیم کرد. روش‌های گروهی شبیه به روش‌های تک رده‌بند عمل می‌کنند، ولی با توجه به استفاده از داده‌های جریان در آموزش رده‌بند و استفاده از رده‌بندهای گروهی مقاومت آن‌ها در برابر نویز بیشتر شده و مفاهیم تکرار شونده را نیز می‌توانند تشخیص دهند. الگوریتم‌های یادگیری کنونی اغلب از روش رده‌بندی گروهی بهره می‌برند.

رده‌بندی فرآیند یافتن یک مدل کلی از داده‌های گذشته برای اعمال بر روی داده‌های جدید است. یک الگوریتم رده‌بندی، این امکان را به ما می‌دهد که به صورت مکرر پیش‌بینی‌هایی با استفاده از پایگاه

[1] Batch learning

اطلاعاتی موجود داشته باشیم. رده‌بندی در دو مرحله آموزش و آزمون انجام می‌شود. در مرحله آموزش، سیستم تلاش می‌کند یک مدل از مجموعه داده‌ها را که مجموعه آموزشی نام دارد، یاد بگیرد. در مرحله آزمون، از مدل برای اختصاص برچسب کلاس برای داده‌های بدون برچسب در مجموعه آزمون استفاده می‌شود. از کاربردهای رده‌بندی می‌توان به بازاریابی، تشخیص بیماری‌های پزشکی، تحلیل داده‌های چند رسانه‌ای، تحلیل داده‌های بیولوژیکی، رده‌بندی و فیلتر اسناد و تحلیل شبکه‌های اجتماعی اشاره کرد [۳۶، ۳۷]. تکنیک‌های مختلفی برای رده‌بندی جریان‌های داده در پژوهش‌ها وجود دارد که می‌توان به درخت تصمیم، رده‌بندی بیزین[۱]، شبکه عصبی، ماشین بردار پشتیبان[۲]، k نزدیک‌ترین همسایه[۳] و رده‌بندهای گروهی[۴] اشاره کرد [۲۲].

۱-۴ هدف پژوهش

هدف اصلی این پایان‌نامه رده‌بندی جریان داده با استفاده از الگوریتم‌های یادگیری گروهی است. در این پایان‌نامه روش جدیدی برای رده‌بندی نیمه‌نظارتی جریان داده ارائه می‌گردد. رده‌بندهای گروهی می‌توانند به روز رسانی شده و تغییرات موجود در جریان داده را پشتیبانی کنند. در این سیستم پس از ساخت مدل‌ها با استفاده از داده‌های آموزشی، برچسب کلاس‌ها برای داده‌های بدون برچسب پیش‌بینی می‌شود. این برچسب‌ها با در نظر گرفتن تغییر مفهوم به داده‌ها اختصاص داده می‌شود. جریان داده تکاملی به جریانی از داده‌ها اطلاق می‌گردد که کلاس واقعی بخشی از داده‌های جریان در مجموعه داده آموزشی وجود نداشته و یادگیر با توجه به ویژگی‌های داده‌ها، کلاس جدید را پیش‌بینی کند.

برای به روز رسانی این رده‌بندها، ابتدا تغییرات موجود در داده‌ها با استفاده از روشی مبتنی بر آنتروپی تشخیص داده شده و در یک بافر نگه‌داری می‌گردد. سپس زمانی که بافر به تعداد مشخصی رسید، مدل

[1] Bayesian classification
[2] Support Vector Machine (SVM)
[۳] K-Nearest neighbor (K-NN)
[۴] Ensemble Classifiers

جدیدی با استفاده از داده‌های درون بافر ساخته شده و به مدل‌های گروهی موجود اضافه می‌شود و این داده‌ها از بافر حذف می‌گردد.

مدل ارائه شده از یک روش یادگیری نیمه‌نظارتی برای رده‌بندی جریان داده استفاده کرده که این روش علاوه بر تشخیص تغییر مفهوم، مفاهیم جدید را نیز شناسایی می‌کند. این مدل با تشخیص سریع تغییرات به خوبی رده‌بند را با تغییرات موجود در جریان داده تطبیق داده و از کاهش دقت رده‌بندی، بدلیل وجود این تغییرات جلوگیری می‌کند. همچنین الگوریتم یاد شده با تشخیص مفاهیم جدید به خوبی دقت رده‌بند را حفظ می‌کند.

۱-۵ مروری بر فصل‌ها

پس از بیان مقدماتی که در این فصل داده شد، در فصل دوم ابتدا مفاهیم مرتبط با موضوع پژوهش ارائه شده و سپس به بررسی روش‌های مرتبطی که تاکنون در زمینه‌های رده‌بندی جریان داده، تغییر مفهوم و جریان داده تکاملی به انجام رسیده است، می‌پردازیم. در فصل سوم به طور کامل راهکار پیشنهادی این پژوهش شرح داده می‌شود. روش پیشنهادی، به همراه جزئیات مورد نیاز برای درک کامل موضوع آورده شده است. در فصل چهارم به بررسی نتایج و بیان نظری و عملی پژوهش پرداخته می‌شود و نمودارهای خروجی حاصل از این پژوهش تحلیل می‌گردد. همچنین نتایج حاصل از روش ارائه شده در این پایان‌نامه با روش‌های پیشین و مرتبط مقایسه و بحث می‌گردد. در نهایت در فصل پنجم جمع‌بندی و نتیجه‌گیری از کل پژوهش به عمل می‌آید و پیشنهاداتی برای توسعه این پژوهش در اختیار خواننده قرار می‌دهد.

فصل ۲ : ادبیات پژوهش

داده کاوی فرآیند استخراج الگو و اطلاعات پنهان از پایگاه داده بزرگ و بکارگیری آن در تصمیم‌گیری‌ها و فعالیت‌های مهم تجاری است. داده کاوی سنتی بیشتر بر روی داده‌های ایستا متمرکز بوده، درحالی‌که امروزه پیشرفت تکنولوژی باعث پدید آمدن جریان‌های داده، تغییر در پردازش داده‌ها و ذخیره‌سازی اطلاعات شده است. اینگونه از داده‌ها که باعث تغییر در روند پردازش داده‌ها شده‌اند، داده‌های پویا می‌نامند. به طور کلی کنترل داده‌های ایستا ساده‌تر بوده و پیچیدگی پردازشی کمتری دارند. امروزه مجموعه داده‌ها برای حافظه‌های اصلی بسیار بزرگ بوده و نیاز به ذخیره‌سازی در حافظه‌های جانبی دارند. بنابراین استفاده از روش‌های دسترسی تصادفی مطابق آنچه که در روش‌های سنتی داده کاوی بکار می‌رفت، بسیار هزینه‌بر است. یکی از اهداف کاوش جریان داده، ساخت مدل با استفاده از نمونه‌های موجود در جریان داده است. البته بازآموزی مدل با استفاده از این نمونه‌ها ناکارآمد بوده و باید به روز رسانی مدل به صورت افزایشی باشد [۱۸].

پیش از این‌که داده کاوی صورت گیرد باید داده‌های مورد نظر شناسایی و اصلاح شوند. برای این کار باید قبل از داده کاوی مراحل پیش‌پردازش انجام شود. این مراحل عبارت است از :

۱. **پاک‌سازی داده:** برای حذف نویز و اغتشاش از داده‌های خام[۱]
۲. **یکپارچه‌سازی داده**: ترکیب داده‌ها از منابع مختلف
۳. **انتخاب داده**: محدود کردن داده به پایگاه‌های داده مرتبط
۴. **تبدیل داده**: تبدیل داده متناسب با نیاز الگوریتم داده‌کاوی
۵. **داده کاوی**: اجرای الگوریتم‌ها برای استخراج الگوها
۶. **ارزیابی الگو**: شناسایی الگوهای جذاب بر اساس معیارهای جذابیت

[۱] Raw data

۷. **ارائه دانش**: ارائه الگوهای یافت شده با تصویرسازی و تکنیک‌های ارائه دانش به کاربر

الگوریتم‌های سنتی داده کاوی به دلیل این که نیاز به چندین گذر بر روی داده‌ها دارند، برای کنترل جریان داده مناسب نیستند. این مسئله یک چالش در حوزه جریان داده بوجود می‌آورد و آن مناسب بودن الگوریتم‌های موجود داده کاوی برای مجموعه داده‌های ایستا و به کار گرفته نشدن برای کنترل و استخراج جریان‌های داده است [۱۸]. امروزه از روش‌های نوینی همچون خلاصه‌سازی داده‌ها، ساز و کار فراموشی داده‌ها و استفاده از پنجره‌گذاری[1] برای مدیریت این چالش استفاده می‌کنند. در ادامه این فصل به بیان و بررسی مفاهیم اساسی مورد نیاز برای انجام پایان نامه پرداخته شده است.

۲-۱ جریان داده

جریان داده در کاربردهای فراوانی ظاهر می‌شود. نمونه‌های فراوانی از جریان‌های داده وجود دارند که می‌توان به داده‌های بدست آمده از حسگرهای بی‌سیم[۲]، اطلاعات حملات شبکه‌ای، بررسی داده‌های ترافیک شبکه، بررسی تراکنش‌های مالی، داده‌های موجود در شبکه‌های اجتماعی و اطلاعات روزانه سبدهای خرید اشاره کرد. از مهمترین چالش‌های موجود می‌توان به برخط بودن، اندازه نامحدود داده‌ها، پیچیدگی در سازمان‌دهی آن‌ها، ماهیت پویا و افزایشی و عدم دسترسی به کل داده‌ها اشاره کرد. از جمله ویژگی‌های جریان داده که آن‌ها را متمایز می‌کند، طول بی‌نهایت، تغییر مفهوم، تکامل مفهوم[3] و تعداد محدود داده‌های برچسب‌دار است [۱۹].

[1] Windowing
[۲] Wireless sensor
[3] Concept evolution

۲-۱-۱ ویژگی‌های جریان داده

ویژگی‌های جریان‌های داده که باید در کاوش مدنظر قرار گیرد [۲،۱]:

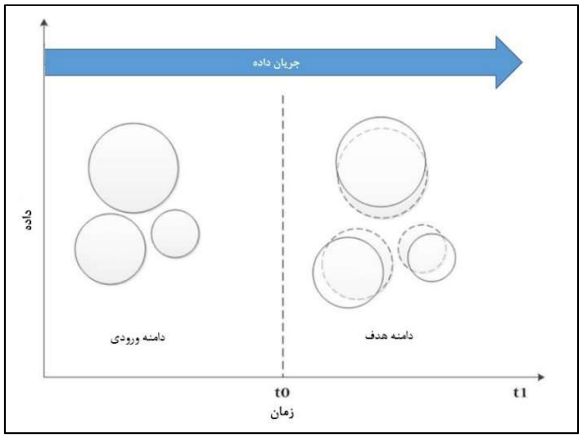
۱. **حجم عظیم جریان داده**: استفاده از الگوریتم‌های متداول یادگیری به‌دلیل نامحدود بودن جریان داده، عدم توانایی ذخیره‌سازی کل داده‌ها و همچنین زمان یادگیری نامحدود آن‌ها، میسر نیست. بنابراین ساز و کاری باید فراهم شود تا یادگیر بر اساس داده‌های مرتبط تصمیم‌گیری کند.

۲. **پردازش برخط**: فرآیند یادگیری در این داده‌ها باید با یک گذر روی داده‌ها انجام شود. از طرفی الگوریتم‌های متداول داده‌کاوی نیاز به چندین گذر روی داده‌ها و دسترسی به داده‌های قدیمی دارند، که به دلیل محدودیت حافظه موجود نیستند و جهت پردازش حجم عظیم جریان داده بسیار کند و غیر عملی هستند.

۳. **وجود تغییر مفهوم**: تغییر مفهوم زمانی اتفاق می‌افتد که توزیع داده‌ها در طول زمان تغییر کند [۱۲]. بنابراین مدل رده‌بند باید با توجه به توزیع جدید به صورت پیوسته به روز گردد.

۲-۱-۲ تغییر مفهوم

تغییر مفهوم به معنای تغییر توزیع ویژگی یا برچسب نمونه‌هاست. در این حالت در داده‌ها کلاس جدیدی پدیدار نمی‌گردد و فقط در توزیع کلاس‌های موجود تغییرات اندکی اتفاق می‌افتد. با شناسایی این تغییرات، می‌توان تصمیمات بهتری در حوزه‌هایی مانند بازار سهام، تغییرات آب و هوایی و رشد جمعیت گرفت. به عنوان مثالی از تغییر مفهوم در دنیای واقعی، می‌توان به کاهش علاقه افراد به یک محصول به دلیل عرضه شدن محصولی با کیفیت بهتر اشاره کرد [۲۰]. در شکل ۲-۱ تغییر مفهوم در طی زمان به خوبی نشان داده شده است.



شکل ۲-۱.تغییر مفهوم [۲۰]

۲-۱-۳ انواع تغییر مفهوم

تغییر مفهوم به چهار دسته ناگهانی[1]، تدریجی[2]، افزایشی[3] و تکرارشونده[4] تقسیم می‌شود [۱۳]. در ادامه با فرض این‌که فقط دو مفهوم داریم به بررسی هر یک از آن‌ها که در شکل ۲-۲ قابل مشاهده هستند، می‌پردازیم:

۱. **تغییر مفهوم ناگهانی**: ساده‌ترین نوع تغییر مفهوم است، به‌طوری‌که در یک زمان مشخص مفهوم جدیدی جایگزین مفهوم قبلی می‌گردد. 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.

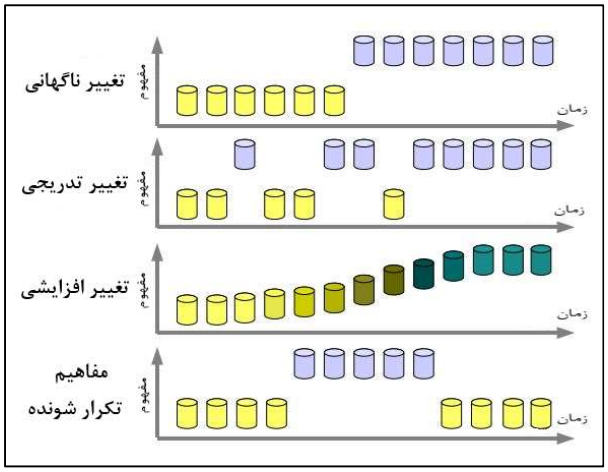
۲. **تغییر مفهوم تدریجی**: در این نوع تغییر مفهوم، مفهوم به تدریج از یک نوع به نوعی دیگر تغییر می‌کند. در یک بازه زمانی هر دو مفهوم فعال هستند، ولی به تدریج از احتمال انتخاب نمونه‌ها برای نوع اول کم شده و به احتمال انتخاب نمونه‌ها برای نوع دوم اضافه می‌شود. برای

[1] Sudden
[2] Gradual
[3] Incremental
[4] Recurring

نمونه زمانی که سرمربی تیم یک کشور تغییر می‌کند، در آن کشور جستجوها در ارتباط با آن شخص به مرور کم شده و برای شخصی که به تازگی سرمربی تیم شده به تدریج 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 به تدریج از نوع اول فاصله گرفته و به نوع دوم نزدیک می‌شود. برای مثال می‌توان به افزایش تدریجی محبوبیت یک نرم‌افزار پس از اضافه شدن قابلیت‌های جدید به آن، اشاره کرد.

۴. **مفهوم تکرار شونده**: در این نوع تغییر مفهوم، مفهومی که در گذشته فعال بوده، می‌تواند بعد از مدت زمانی دوباره پدیدار شود. مثلا فرض کنید کاربر یک کشاورز است که در فصل بهار سمومی که باید در این فصل به درختان بزند را در اینترنت جستجو می‌کند. این کاربر در ابتدای بهار هر سال این اطلاعات را دوباره جستجو می‌کند.

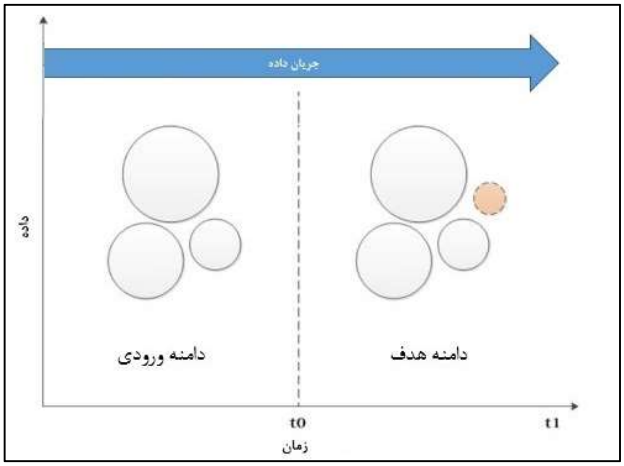


شکل ۲-۲.انواع تغییر مفهوم [۱۳]

۲-۱-۴ تکامل مفهوم

زمانی که تغییر داده‌ها موجب پدیدار شدن مفاهیم جدید می‌گردد، تکامل مفهوم اتفاق افتاده و به صورت ظاهر شدن کلاس جدید در حین تکامل نمایان می‌شود. با بررسی داده‌های موجود در جریان داده، دنباله‌ای از داده‌ها که از تمام مفاهیم موجود دور شده و یک شکل و قالب پایدار و مستقل به خود می‌گیرند به عنوان یک مفهوم جدید[1] در نظر می‌گیریم [۲۰]. در شکل ۲-۳ می‌توانید تکامل مفهوم را مشاهده نمایید.

همانطور که مشاهده می‌کنید تفاوت تکامل مفهوم با تغییر مفهوم در فاصله داده‌ها از مفاهیم موجود و پایداری آن داده‌ها است. در تکامل مفهوم برای تشکیل کلاس جدید باید داده‌ها به اندازه کافی از سایر کلاس‌های موجود دور بوده و نسبت به هم نزدیک باشند و همچنین باید از نظر تعداد به یک حد نصاب برسند. ولی در تغییر مفهوم داده باید به یکی از کلاس‌های موجود نزدیک بوده و نیازی به حد نصاب از نظر تعداد هم ندارند.

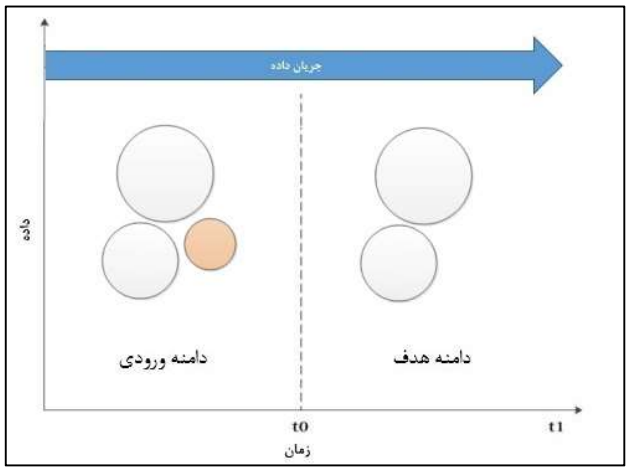


شکل ۲-۳.تکامل مفهوم [۲۰]

[1] Novel concept

۲-۱-۵ فراموشی مفهوم[1]

زمانی که تعداد داده اندکی در طول زمان به یک مفهوم اختصاص یابد، به این معنی است که مفهوم مورد نظر در حال ناپدید شدن است. در چنین شرایطی که کلاس مربوط به آن مفهوم تغییری نداشته باشد، مفهوم مورد نظر به فراموشی سپرده می‌شود [۲۰]. در شکل ۲-۴ فراموشی مفهوم در طول زمان نشان داده شده است.

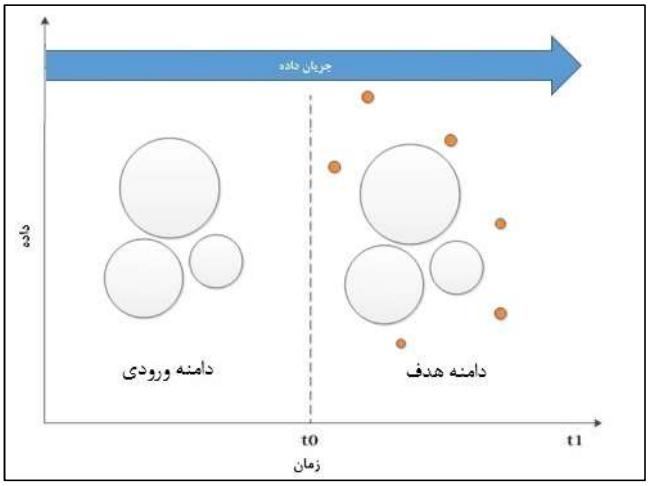


شکل ۲-۴.فراموشی مفهوم [۲۰]

۲-۱-۶ داده‌های دور افتاده[2]

داده دور افتاده به داده‌ای گفته می‌شود که نسبت به مفاهیم موجود و اکثر داده‌ها فاصله زیادی داشته و همچنین نسبت به یکدیگر نزدیک نباشند [۲۰]. بنابراین نه قابلیت داده‌های تغییر مفهوم را دارند که بتوانند به کلاس‌های موجود اضافه شوند و نه قابلیت داده‌های تکامل مفهوم را دارند که بتوانند کلاس جدیدی تشکیل دهند. در شکل ۲-۵ می‌توان داده‌های دور افتاده را مشاهده کرد.

[1] Concept forgetting
[2] Outlier



شکل ۲-۵.داده‌های دور افتاده [۲۰]

۲-۱-۷ نرم‌افزارهای کار با جریان داده

از جمله نرم‌افزارهای متن بازی[1] که برای کار بر روی جریان داده طراحی شده‌اند، می‌توان به چند مورد زیر اشاره کرد:

۱. **WEKA** : شناخته شده ترین نرم‌افزار داده کاوی در محیط علمی است. WEKA شامل مجموعه‌ای از الگوریتم‌های یادگیری مانند پیش پردازش اطلاعات، رده‌بندی، رگرسیون، خوشه‌بندی و بصری سازی است [۱۷].

۲. **MOA**[۲] : یک محیط نرم‌افزاری بر پایه WEKA است که برای اجرای الگوریتم‌ها و اجرای آزمایشات در زمینه یادگیری برخط جریان داده به کار برده می‌شود. MOA شامل بسیاری از الگوریتم‌های یادگیری برخط برای تکامل جریان داده مانند درخت تصمیم‌گیری خیلی سریع[3]

[1] Open source
[۲] Massive Online Analysis
[3] Very Fast Decision Tree

و الگوریتم‌های یادگیری گروهی می‌باشد. علاوه بر این MOA مولدهای ساخت جریان داده از قبیل SEA Concept، STAGGER، Random RBF Generator و Random RBF Generator Drift را فراهم می‌کند [۱۶].

۳. **RapidMiner** : یکی دیگر از نرم‌افزارهای کار بر روی جریان داده است که از WEKA قدرتمندتر بوده و علاوه بر الگوریتم‌های موجود در آن، الگوریتم‌های پیشرفته دیگری نیز در اختیار دارد [۱۷].

۴. **Spark** : یکی از سریع‌ترین و شناخته شده‌ترین نرم‌افزارهای پردازش داده‌های حجیم می‌باشد. این ابزار از حافظه اصلی برای نگهداری داده‌ها کرده و به همین دلیل برنامه‌ها سریع‌تر اجرا می‌شوند. همچنین یکی دیگر از مواردی که موجب افزایش کارایی آن می‌شود، استفاده از مکانیزم حافظه نهان است که داده‌هایی که دوباره در برنامه قرار است اجرا شوند، با استفاده از این مکانیزم ذخیره می‌شوند [۴۲].

۲-۲ رده‌بندی

اصطلاح رده‌بندی مربوط به هر زمینه‌ای است که در آن تصمیمی گرفته شده و یا اینکه بر اساس یکسری اطلاعات موجود، پیش‌بینی صورت گرفته باشد. یک الگوریتم رده‌بندی، این امکان را به ما می‌دهد که به صورت مکرر پیش‌بینی‌هایی با استفاده از پایگاه اطلاعاتی موجود داشته باشیم. پس از آن‌که یک الگوریتم رده‌بندی بر اساس اطلاعات موجود ساخته می‌شود؛ از آن برای رده‌بندی مواردی که مشاهده نشده، اعمال می‌شود. از کاربردهای رده‌بندی می‌توان به بازاریابی، تشخیص بیماری‌های پزشکی، تحلیل داده‌های چندرسانه‌ای، تحلیل داده‌های بیولوژیکی، رده‌بندی و فیلتر اسناد و تحلیل شبکه‌های اجتماعی اشاره کرد [۳۶، ۳۷]. رده‌بندی در زمینه داده‌کاوی فرآیند یافتن یک مدل کلی از داده‌های گذشته برای اعمال بر روی داده‌های جدید است. فرآیند رده‌بندی شامل دو مرحله آموزش و آزمون است. در مرحله

آموزش، سیستم بر روی یک مدل از مجموعه داده‌ها که مجموعه آموزشی نام دارد، آموزش می‌بیند. در مرحله آزمون، از مدل برای اختصاص برچسب کلاس به داده‌های بدون برچسب در مجموعه آزمون استفاده می‌شود. تکنیک‌های مختلفی برای رده‌بندی جریان داده در پژوهش‌ها وجود دارد که می‌توان به درخت تصمیم، رده‌بندی بیزین، شبکه عصبی، ماشین بردار پشتیبان، k نزدیک‌ترین همسایه و رده‌بندهای گروهی اشاره کرد [۲۲].

فرآیند رده‌بندی با خوشه‌بندی متفاوت است. به این دلیل که در رده‌بندی برچسب نمونه‌های ورودی از قبل مشخص است، در حالی که در خوشه‌بندی نمونه‌های ورودی برچسب اولیه ندارند.

۲-۲-۱ رده‌بندی جریان داده

با توجه به فراوانی جریان داده در کاربردهای روزمره، رده‌بندی جریان داده از حوزه‌های مهم و پرکاربرد به شمار می‌رود. در همین راستا، نیاز به درک مفاهیم مربوط به چگونگی برخورد و کار با این‌گونه داده‌ها تا حد زیادی احساس می‌شود. فرآیند رده‌بندی بر روی جریانی از داده‌ها با چالش‌هایی از قبیل اعمال محدودیت بر روی داده‌ها و تغییرات دائمی و با سرعت زیاد داده‌ها مواجه است. به لحاظ تئوری، طول یک جریان داده بی‌نهایت است. برای کمک به این موضوع که کدام بخش از جریان داده به الگوهای داده کاوی کمک می‌کند، مدل‌های پنجره مورد استفاده قرار می‌گیرند. ۴ نوع مدل پنجره وجود دارد که در زیر به آن‌ها اشاره شده است [۲۱،۲۲]:

- **پنجره‌گذاری از یک نقطه مشخص**[۱] : در این مدل، رده‌بندی از یک نقطه شروع آغاز شده و تا لحظه مورد نظر ادامه پیدا می‌کند. برای این پنجره می‌توان تعدادی نمونه مشخص و یا یک بازه زمانی تعیین کرد. در این مدل، زمانی که یک پنجره جدید از داده‌ها شروع می‌شود؛ تمامی

[۱] LandMark window

نمونه‌هایی که از قبل نگهداری شده‌اند، حذف می‌شوند. همچنین به داده‌های موجود در پنجره وزن یکسانی داده شده و اهمیت داده‌ها یکسان هستند.

- **پنجره‌های لغزان**[۱] : در این مدل، پنجره‌ای با اندازه ثابت وجود داشته و رده‌بندی از یک نقطه شروع مشخص آغاز می‌شود. با گذشت زمان، طول این پنجره ثابت بوده و روی داده‌ها حرکت می‌کند. این مدل بیشتر برای کاربردهایی مناسب است که در آن‌ها جدیدترین نمونه‌های داده از اهمیت بیشتری برخوردار باشند.

- **پنجره‌های محو شونده**[۲] : در این مدل، به هر یک از نمونه‌های جریان داده بر اساس زمان ورود یک وزن اختصاص داده می‌شود. زمانی که یک نمونه جدید وارد می‌شود، بیشترین وزن ممکن به آن نمونه اختصاص داده می‌شود. با گذشت زمان وزن تخصیص داده شده، کاهش می‌یابد. در این مدل بر خلاف دو نوع قبل، نمونه‌ها به طور کامل حذف نشده و فقط با کم کردن وزن داده‌ها، تاثیر استفاده از آن‌ها را کم می‌کند.

- **پنجره متمایل**[۳] : در این مدل، با توجه به زمان ورود داده‌ها سطوح مختلفی بر روی آن‌ها اعمال می‌شود. برای داده‌های اخیر مقیاس کوچکتر و برای داده‌های گذشته مقیاس بزرگتری در نظر گرفته می‌شود. این مدل تقریبا تمام داده‌ها را ذخیره کرده و تناسب خوبی بین دقت و فضای ذخیره‌سازی مورد نیاز برقرار می‌کند. این مدل ممکن است، بعد از اجرای طولانی مدت ناپایدار شود. برای مثال ساختار درخت FP-Stream به گونه‌ای است که در طول زمان بسیار بزرگ شده و فرآیند به روز رسانی و پویش کل درخت، عملکرد این مدل را پایین می‌آورد [۲۳].

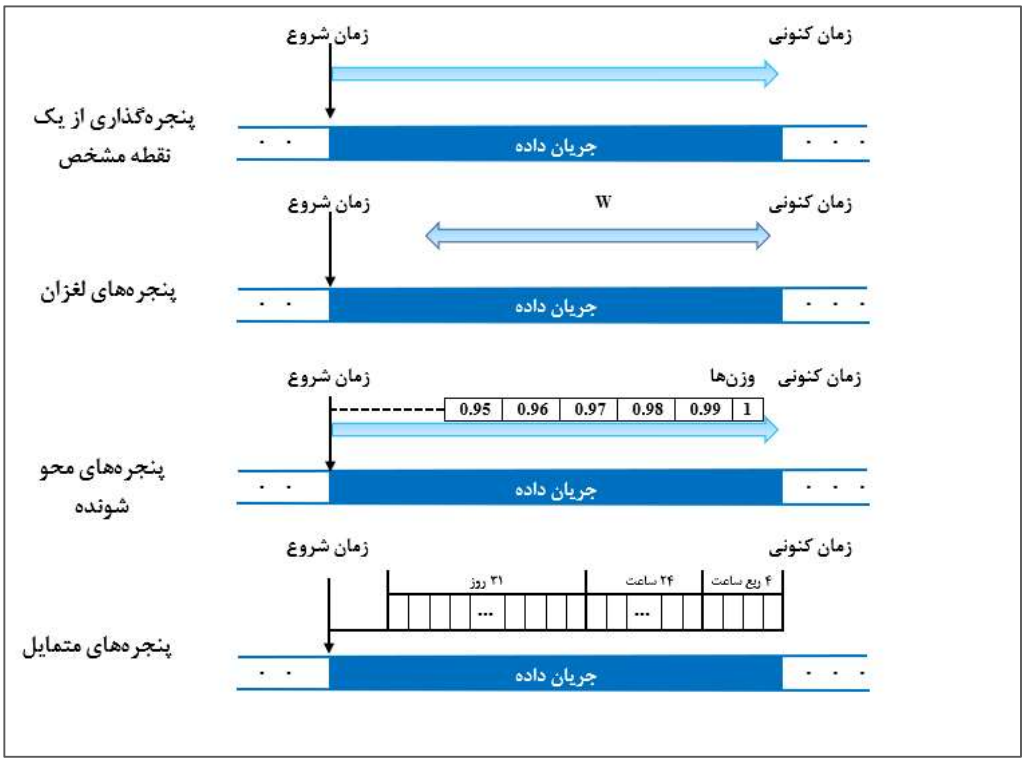
---

۱ Sliding window

۲ Fading window

۳ Tilted time window

به منظور روشن‌تر شدن تعاریف مربوط به مدل‌های پنجره‌ای مطرح شده برای کار با جریان داده‌ها شکل ۲-۶ در ادامه قرار گرفته است. این شکل نمونه‌هایی از چهار مدل پنجره را نشان می‌دهد. برای پنجره متمایل چهار ربع آخر از یک ساعت، آخرین ۲۴ ساعت و در نهایت ۳۱ روز آخر ذخیره شده است.



شکل ۲-۶.نمایش مدل‌های مختلف پنجره‌گذاری [۱۷]

۲-۲-۲ انواع روش‌های رده‌بندی جریان داده

تاکنون روش‌های متعددی برای رده‌بندی جریان داده معرفی شده است که در این بخش به بررسی چند مورد از روش‌های مطرح موجود در این زمینه می‌پردازیم.

۱. **درخت تصمیم** : درخت تصمیم یکی از قوی‌ترین و پرکاربردترین الگوریتم‌های رده‌بندی جریان داده است. درخت هافدینگ[۱] یک رده‌بند درخت تصمیم برای جریان داده است. درخت‌های تصمیم سنتی نیاز به چندین بار پویش بر روی داده‌های آموزشی برای جداسازی ویژگی‌ها دارند. در حالی که در جریان داده این امر امکان پذیر نبوده و در چنین محیطی غیر قابل اجرا هستند. برای حل این مشکل درخت هافدینگ از یک حد آستانه برای انتخاب ویژگی مربوط به جداسازی استفاده می‌کند. الگوریتم درخت هافدینگ یک الگوریتم افزایشی است [۲۴]. برای ورود هر داده جدید، از حد آستانه برای انتخاب بهترین ویژگی مربوط به جداسازی استفاده کرده و بررسی می‌کند که ویژگی انتخاب شده به اندازه کافی برای ساخت سطح بعدی درخت مناسب هست یا خیر. این الگوریتم به دلیل عدم تغییر در گره[۲] ساخته شده، قادر به تشخیص تغییر مفهوم در جریان داده نیست. الگوریتم درخت تصمیم‌گیری خیلی سریع سازگار با مفهوم[۳] توسعه یافته الگوریتم درخت هافدینگ بوده و قادر به تشخیص تغییر مفهوم می‌باشد [۲۵]. این الگوریتم در هر گره برای نظارت بر اعتبار تصمیمات گرفته شده، مقداری را به آن گره نسبت می‌دهد. با ورود داده‌ها، به صورت پیوسته مقادیر ذخیره شده در گره‌ها به روز رسانی می‌شوند. با کم شدن مقادیر گره‌های متناظر با داده‌های قدیمی، این داده‌ها حذف می‌شوند. با پویش دوره‌ای گره‌ها، الگوریتم درخت تصمیم‌گیری خیلی سریع سازگار با مفهوم قادر به تشخیص تغییر مفهوم می‌گردد. در صورتی که تغییر تشخیص داده شود؛ با بزرگ کردن شاخه‌ای که بهترین ویژگی را دارد، قدیمی‌ترین شاخه حذف می‌شود.

۲. **رده‌بندی بیزین** : الگوریتم مدل مقایسه‌ای بیزین[۴] از قضیه احتمالاتی برای اختصاص درجه باور به مدل‌های مد نظر با توجه به مشاهدات و باورهای قبلی استفاده می‌کند. در (رابطه ۲-۱)

---

[۱] Hoeffding Tree

[۲] Node

[۳] Concept-Adaptive Very Fast Decision Tree(CVFDT)

[۴] Bayesian Model Comparison (BMC)

قضیه بیز نمایش داده شده که M مجموعه‌ای از مدل‌های مورد بررسی بوده و D مجموعه مشاهدات است. این قضیه از روشی بر پایه احتمال وقوع یا عدم وقوع برای رده‌بندی استفاده می‌کند و احتمال رخ دادن یک پدیده محاسبه و رده‌بندی صورت می‌گیرد.

$$p(M|D) = \frac{p(D|M)p(M)}{p(D)} \quad \text{(رابطه ۲-۱)}$$

پژوهش‌هایی که در زمینه بیزین انجام گرفته‌اند، از مدل مقایسه‌ای بیزین برای تشخیص تغییر در داده‌ها استفاده کرده‌اند. برای رسیدن به این هدف دنباله جریان داده به پنجره‌هایی با توزیع یکسان تقسیم شده و پنجره‌ها به صورت مستقل از هم فرض می‌شوند [۲۶]. برای نمونه فرن‌هد و همکارش [۲۷] روشی پیشنهاد کرده‌اند که با استفاده از توزیع پسین[۱] در هر یک از کلاس‌ها تغییرات را شناسایی می‌کند.

۳. **شبکه عصبی** : شبکه عصبی از پرکاربردترین و عملی‌ترین روش‌های مدل‌سازی مسائل پیچیده و بزرگ است. هر شبکه عصبی شامل یک لایه ورودی است که هر گره در این لایه معادل یک از متغیرهای پیش‌بینی می‌باشد. قاضی‌خانی و همکارانش [۲۸] یک شبکه عصبی برخط پیشنهاد کردند که شامل دو بخش کنترل تغییر مفهوم و عدم توازن کلاس‌ها است. تغییر مفهوم با استفاده از یک مکانیزم مبتنی بر فراموشی شناسایی شده و وظیفه آن وزن‌دهی به نمونه‌های اخیر جریان داده است. مشکل عدم توازن کلاس‌ها هم با استفاده از یک تابع خطا و پرسپترون چند لایه که دارای تابع فعال‌ساز تانژانت هیپربولیک است، برطرف شده است.

۴. **ماشین بردار پشتیبان** : ماشین‌های بردار پشتیبان عملکرد برجسته‌ای در بسیاری از مشکلات مربوط به یادگیری ماشین با مجموعه داده‌های ایستا دارند. با این حال پیچیدگی زمانی و حافظه آن‌ها بالاست. تسانگ و همکارانش [۲۹] الگوریتم ماشین بردار هسته[۲] را ارائه کرده‌اند

[۱] Posterior

[۲] Core Vector Machine

که این پیچیدگی‌ها را کاهش می‌دهد. ابتدا الگوریتم یک فضایی از مجموعه داده که با تقریب مناسبی مشابه با مجموعه داده اصلی است، پیدا می‌کند. سپس بر روی این فضا بیشترین فاصله را می‌یابد. رای و همکارانش [۳۰] الگوریتم StreamSVM را که توسعه یافته ماشین بردار هسته است، ارائه کرده‌اند. این الگوریتم می‌تواند با یکبار پویش بر روی جریان داده کار رده‌بندی را انجام دهد. در این الگوریتم شعاع فضای در نظر گرفته شده متغیر بوده و زمانی که داده‌های آموزشی اضافه می‌شود، شعاع زیاد می‌شود.

۵. **k نزدیک‌ترین همسایه** : قاعده k نزدیکترین همسایه در واقع رده‌بندی یک داده بر اساس رای اکثریت بر روی کلاس k نزدیکترین همسایه داده مورد نظر است که کلاس این داده‌های همسایه از قبل مشخص شده‌اند. آگاروال و همکارانش [۳۱] یک الگوریتم برای رده‌بندی جریان داده مبتنی بر K-NN ارائه کرده‌اند که توسعه‌ای بر روش Clustream است. ساختار این الگوریتم میکرو خوشه بوده و دارای دو رویکرد برخط و برون خط[۱] می‌باشد. در فرآیند رده‌بندی برون خط به دنبال بهترین پنجره داده بوده و میکرو خوشه‌ها استخراج می‌شوند. در رده‌بندی 1-NN با اختصاص داده‌های آزمون به برچسب نزدیک‌ترین میکرو خوشه، کلاس داده آزمون تعیین می‌گردد.

۶. **رده‌بندهای گروهی** : Bagging و Boosting دو رده‌بند گروهی هستند که عملکرد بهتری نسبت به سایر الگوریتم‌ها در مجموعه داده‌های سنتی دارند. بنابراین بسیاری از پژوهشگران تلاش کرده‌اند این دو روش را بر روی جریان داده تطبیق دهند. اوزا [۳۲] یک روش بر پایه این دو الگوریتم ارائه کرده‌اند. با استفاده از (رابطه ۲-۲) احتمال متناسب با هر داده آموزشی که k بار در رده‌بند گروهی ظاهر شده برابر :

[۱] Offline

$$P(k) = \binom{N}{k}\left(\frac{1}{N}\right)^{k}\left(1-\frac{1}{N}\right)^{N-k}$$ (رابطه ۲-۲)

است که k اندازه مجموعه داده آموزشی و N اندازه کل مجموعه داده می‌باشد. در جریان داده فرض می‌شود، تعداد داده‌ها بی‌نهایت است. بنابراین احتمال $P(k)$ به (رابطه ۲-۳) تغییر می‌کند.

$$P(k) = \frac{e^{-1}}{k!}$$ (رابطه ۲-۳)

در این روش با استفاده از الگوریتم Bagging به هر یک از داده‌ها وزنی مطابق با احتمال بدست آمده از (رابطه ۲-۳) داده شده و در الگوریتم Boosting با توجه به میزان خطای رده‌بند وزن داده‌ها تنظیم می‌گردد.

۲-۳ یادگیری نیمه‌نظارتی

روش‌های یادگیری را می‌توان با توجه به میزان داده‌های آموزشی برچسب‌دار به سه دسته با نظارت، نیمه‌نظارتی و بدون نظارت 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 درصورتی که علاوه بر داده‌های برچسب‌دار، از داده‌های بدون برچسب در یادگیری استفاده شود، یادگیری نیمه‌نظارتی خواهد بود. برای جریان‌های داده، دسته نیمه‌نظارتی الگوریتم‌های یادگیری بیش از دسته نظارتی مورد نیاز است؛ چرا که همانطور که عنوان شد، در جریان‌های داده فرض وجود برچسب برای همه داده‌ها یک فرض غیر واقع‌بینانه بوده و تعداد داده‌های برچسب‌دار نسبت به کل داده‌ها اندک است [۱۱]. در بسیاری از برنامه‌های کاربردی عملیات برچسب‌گذاری هزینه‌بر و زمان‌بر بوده و برای افزایش کارایی و بهبود دقت می‌توان از داده‌های بدون برچسب استفاده کرد. به این صورت که با استفاده از داده‌های برچسب‌دار مدلی از رده‌بند ساخته و هر نمونه بدون برچسب را بر اساس مدل‌های ساخته

شده، به کلاس شبیه‌تر اختصاص دهیم. در یادگیری نیمه‌نظارتی هدف یافتن راهکارهایی است که بتوان با استفاده از آن به حل مسئله پرداخته و در نتیجه کارایی الگوریتم افزایش یابد.

به طور کلی در یادگیری نیمه‌نظارتی کلاس نمونه‌های بدون برچسب با توجه به نمونه‌های برچسب‌دار تعیین شده و این کار موجب ایجاد مطلوب‌تر رده‌بند می‌گردد. الگوریتم‌های نیمه‌نظارتی موجود اغلب در چهار مدل شناخته شده می‌گنجند. در ادامه این چهار مدل را به اختصار بیان می‌کنیم.

۱. **مدل خودآموز**[۱] : در این مدل، رده‌بند به کمک نمونه‌های برچسب خورده آموزش می‌بینند. سپس رده‌بند به منظور پیش‌بینی کلاس نمونه‌های بدون برچسب اعمال می‌گردد. در این روش مجموعه‌ای از قابل اطمینان‌ترین نمونه‌های بدون برچسب با برچسب‌های پیش‌بینی شده انتخاب و سپس به مجموعه آموزشی اضافه خواهند شد. رده‌بند مجددا بر روی مجموعه نمونه‌های جدید آموزش دیده و این رویه تکرار می‌گردد [۱۱].

۲. **مدل آموزش مشترک**[۲] : در این روش یادگیری با استفاده از دو رده‌بند مختلف با دو مجموعه ویژگی متفاوت از نمونه‌های برچسب خورده، صورت می‌گیرد. هر رده‌بند به کمک مجموعه ویژگی‌ها آموزش دیده و برای رده‌بندی نمونه‌های بدون برچسب مورد استفاده قرار می‌گیرد. به طور کلی مدل آموزش مشترک نیاز به دو نما از داده‌ها دارد؛ یعنی هر نمونه، دو مجموعه از ویژگی‌هایی که مختلف و مکمل یکدیگر هستند را برای توصیف یک واقعیت ارائه می‌دهد. داده‌هایی که توسط هر کدام از این رده‌بندها با اطمینان بالایی رده‌بندی شده‌اند، به رده‌بند دیگر داده می‌شوند تا خود را به روز نمایند. آموزش مشترک در سنجش آلودگی هوا و رده‌بندی رایانامه‌ها به کار گرفته شده است [۳۴].

---

۱ Self-training model

۲ Co-training model

۳. **مدل‌های ترکیبی مولد**[۱] : در روش‌های مولد ابتدا یک مدل پارامتری برای تابع توزیع نقاط (مثلا توزیع گاوسی) انتخاب می‌شود که آن را با $P(x|y,\theta)$ نشان می‌دهیم و θ نمایانگر مدل است. سپس $P(y)$ از روی داده‌های برچسب‌دار تخمین زده می‌شود. احتمال وقوع نقاط با توجه به تابع توزیع هر دسته، بر حسب پارامترهای مدل، به صورت تحلیلی محاسبه می‌شود. سپس با اعمال قانون بیز می‌توان تابع توزیع برچسب در هر نقطه را محاسبه کرد. در روش‌های مولد معمولاً هدف بیشینه کردن این احتمال وقوع یا به طور معادل درست‌نمایی بیشینه[۲] آن‌ها نسبت به پارامترهای مدل است. از روش‌های مختلفی می‌توان برای بهینه‌کردن پارامترهای مدل نسبت به میزان راستی‌آزمایی استفاده کرد. در مقابل روش‌های مولد، روش‌هایی نیز وجود دارند که به طور مستقیم به یادگیری $P(y|x)$ می‌پردازند، این روش‌ها با نام روش‌های تمایزی[۳] شناخته می‌شوند. به طور کلی روش‌های مولد نیاز به داده‌های بیشتری نسبت به روش‌های تمایزی جهت یادگیری دارند [۳۴].

۴. **مدل‌های مبتنی بر گراف**[۴] : در این مدل‌ها یک گراف روی داده‌ها ساخته می‌شود. در آن رئوس نمایانگر داده‌ها و یال‌های بدون جهت که می‌توانند وزن‌دار باشند، نمایانگر شباهت[۵] نمونه‌ها هستند. به طور کلی این روش در صورتی موثر است که پیش‌فرض همواری[۶] و خمینه[۷]، هم‌زمان برقرار باشند (لازم به ذکر است که پیش‌فرض‌های عنوان شده در انتهای این بخش بررسی می‌گردد). در روش‌های نیمه‌نظارتی مبتنی بر گراف، ابتدا گراف همسایگی روی نقاط

---

۱ Generative mixture models
2 Maximum likelihood
3 Discriminative
۴ Graph-based models
۵ Similarity
۶ Smoothness assumption
۷ Manifold assumption

ساخته می‌شود و سپس به کمک روش‌های استنتاجی نسبت به تعیین برچسب نقاط بدون برچسب اقدام می‌گردد [۳۵].

به طور کلی زمانی می‌توان از یادگیری نیمه‌نظارتی استفاده کرد که توزیع داده‌های بدون برچسب به مسئله رده‌بندی مربوط باشد. یعنی داشتن اطلاعاتی مربوط به $P(x)$ بتواند در تخمین $P(y|x)$ موثر باشد. معمولا در یادگیری نیمه‌نظارتی پیش‌فرض‌هایی در رابطه با داده‌ها در نظر گرفته می‌شود که شامل موارد زیر خواهد بود [۱۱]:

- **پیش‌فرض همواری** : اگر دو نقطه x_1 و x_2 در یک ناحیه با چگالی بالا نزدیک به هم باشند، برچسب‌های متناظر آن‌ها یعنی y_1 و y_2 نیز باید به یکدیگر نزدیک باشند. به عبارت دیگر مرز تصمیم‌گیری مناطق کم تراکم، نقاط نزدیک به هم کمتری خواهد داشت و نقاط دارای کلاس‌های متفاوتی خواهند بود.

- **پیش‌فرض خوشه**[۱] : اگر نقاط در یک خوشه باشند؛ احتمال اینکه برچسب آن‌ها یکسان باشد، زیاد است. در واقع می‌توان یک الگوریتم خوشه‌بندی روی داده‌ها اجرا کرد، سپس به هر خوشه یک برچسب نسبت داد. طبق این پیش‌فرض می‌توان به این نتیجه رسید که احتمال دیدن داده‌های دسته‌های متفاوت در یک خوشه کم است.

- **پیش‌فرض خمینه** : در فضای ورودی با بُعد بالا، داده‌ها روی یک خمینه با بُعد پایین‌تری قرار دارند و تابع جداساز روی خمینه داده‌ها هموار است. در این وضعیت می‌توان معیار فاصله‌ای مبتنی بر خمینه برای داده‌ها تعریف کرد و داده‌های بدون برچسب را نیز در یادگیری دخالت داد. در این صورت داده‌هایی که طبق این معیار فاصله جدید به یکدیگر نزدیک‌تر هستند، با

[۱] Cluster assumption

احتمال بالاتری برچسب یکسان دارند.این احتمال با افزایش تعداد نقاطی که بین این دو نقطه بر روی خمینه قرار می‌گیرند، افزایش می‌یابد.

۲-۴ بررسی پژوهش‌های انجام شده

در یادگیری دسته‌ای با میانگین‌گیری از مدل‌های تصادفی حاصل از آموزش داده‌ها، بر مشکل محدود بودن داده‌های آموزشی غلبه می‌شود [۴۱]. اما این روش به حجم داده‌ها وابسته بوده و با افزایش حجم داده‌ها از نظر زمان اجرا برای تولید مدل‌ها و تحلیل و بررسی آن‌ها دچار مشکل می‌شود. امروزه اکثر پژوهش‌های رده‌بندی جریان داده با هدف رسیدن به کارایی مطلوب و به حداقل رساندن خطا انجام می‌شوند.

بسیاری از پژوهش‌های انجام شده در زمینه تشخیص تغییر مفهوم بر اساس تخمین توزیع نمونه‌ها بوده و با توجه به این توزیع، تغییر مفهوم را تشخیص می‌دهند [۳۳]. مشکلی که این روش‌ها دارند، وابستگی به نوع داده ورودی است. روش‌های تشخیص تغییر مفهوم را می‌توان بر اساس تعداد رده‌بند، به روش‌های تک رده‌بند یا گروهی تقسیم کرد.

در روش‌های تک رده‌بند، خطای یک رده‌بند روی پنجره داده‌های مختلف محاسبه می‌شود و با توجه به تفاوت خطای رده‌بند روی پنجره‌ها تغییر مفهوم تشخیص داده می‌شود. به عنوان مثال نیشیدا و همکارش [۳۸] از یک آزمون آماری که دو مقدار خطا پیش‌بینی کنونی و کلی رده‌بند را مقایسه می‌کند، استفاده کرده‌اند. در این روش درصورتی که نسبت خطای رده‌بند در پنجره داده اخیر به خطای کل زیاد باشد، تغییر مفهوم تشخیص داده می‌شود.

روش‌های گروهی شبیه به روش‌های تک رده‌بند عمل می‌کنند، ولی با توجه به استفاده از رده‌بندهای گروهی مقاومت آن‌ها در برابر نویز بیشتر شده و مفاهیم تکرار شونده را نیز می‌توانند تشخیص دهند. به

عنوان مثال نیشیدا و همکارانش [۳۹] از روشی مبتنی بر رده‌بندی گروهی سازگار با تغییر مفهوم[۱] استفاده کرده‌اند که خطای همه رده‌بندها را روی پنجره داده اخیر محاسبه می‌کند. در صورتی که هیچ‌یک از این مقادیر خطا قابل قبول نباشد، تغییر مفهوم تشخیص داده می‌شود. مشکل این روش حافظه مصرفی زیاد آن است.

برخی از مقالات از وزن‌دهی برای تشخیص تغییر مفهوم استفاده می‌کنند. برای نمونه سیدو و همکارش [۸] یک روش مبتنی بر وزن‌دهی پویا برای رده‌بندی داده‌ها ارائه می‌دهد. این وزن‌دهی بر پایه دو معیار دقت رده‌بند و پیش‌بینی نهایی الگوریتم است.

برخی از پژوهش‌ها از تکنیک تشخیص تغییر صریح[۲] برای شناسایی تغییر مفهوم در جریان داده استفاده می‌کنند. در کاوش جریان داده، این تکنیک یا برای تشخیص تغییر در توزیع داده‌ها و یا برای تشخیص تغییر در بازخورد رده‌بند استفاده می‌شود. برای نمونه کنچوا و همکارش [۷] تغییر توزیع داده ورودی در یک جریان داده را تشخیص می‌دهد. در پژوهشی دیگر بیفت و همکارش [۵] با استفاده از یک روش مبتنی بر پنجره کشویی، اندازه پنجره را با توجه به میزان تغییرات موجود در داده‌های پنجره مشخص می‌کند. در این روش زمانی یک تغییر تشخیص داده می‌شود که میزان خطا در کل پنجره جاری به طور قابل ملاحظه‌ای از کمترین میزان خطای ثبت شده، بیشتر باشد. مشکل روش‌های فوق آن است که برای شناسایی تغییرات نیاز به برچسب درست برای تمام نمونه‌ها دارند. همچنین فرض مورد استفاده این روش‌ها این است که تغییر مفهوم منجر به کاهش کارایی الگوریتم‌های رده‌بندی می‌شود.

هک و همکارانش [۶] نیز از یک تکنیک تشخیص صریح برای تغییر مفهوم استفاده می‌کند، با این تفاوت که تشخیص تغییر در اطمینان رده‌بند انجام می‌گیرد و با توجه به این که تشخیص بر روی میزان

[۱] Adaptive classifiers ensemble system for concept drifting (ACE)

[2] Explicit concept drift technique

خطای رده‌بند نیست، نیازی به برچسب درست برای تمام داده‌ها نداشته و از یک روش نیمه‌نظارتی بهره می‌برد. با ورود هر نمونه از داده‌های جریان، همراه با پیش‌بینی برچسب نمونه، میزان اطمینان پیش‌بینی را برآورد می‌کند. پس از اضافه شدن مقدار اطمینان به یک پنجره، تشخیص تغییر با توجه به تغییر توزیع مقادیر موجود در پنجره بررسی می‌گردد. در صورتی که تغییر قابل توجهی در برآورد اطمینان مشاهده شود، تغییر تشخیص داده شده و رده‌بند برای انطباق با تغییر باید به روز گردد. این الگوریتم برای داده‌هایی که پیش‌بینی اطمینان ضعیفی داشته، درخواست برچسب درست می‌کند و برای سایر داده‌ها از برچسب پیش‌بینی شده به عنوان برچسب نهایی استفاده می‌کند. در نهایت یک مدل جدید بر روی داده‌های پنجره ساخته شده و رده‌بند گروهی به روز می‌گردد و داده‌های پنجره پاک می‌شود. اگر نمرات اطمینان تغییر قابل توجهی پیدا نکند، مدل گروهی تغییری نکرده و پنجره بزرگ می‌شود. همچنین این مقاله کلاس جدید را با شناسایی داده‌های دور افتاده‌ای که همبستگی قوی بین خود دارند تشخیص می‌دهد. مشکل این روش آن است که برای به روز رسانی مدل‌ها از روش یادگیری فعال استفاده کرده و برچسب تعدادی از نمونه‌ها را از کاربر درخواست می‌کند.

مهدی و همکارانش [۹] از روشی مبتنی بر آنتروپی برای تشخیص تغییر مفهوم استفاده کرده و از یک روش نیمه‌نظارتی برای یادگیری الگوریتم بهره می‌برد. ابتدا دو پنجره با طول ثابت در نظر گرفته و برای هر یک از داده‌های جریان، بر مبنای احتمال وقوع نمونه مقادیری بدست می‌آورد و به پنجره اضافه می‌کند. پس از پر شدن پنجره‌ها، آنتروپی هر کدام از پنجره‌ها محاسبه شده و نمونه‌های دو پنجره با هم ترکیب می‌شوند و آنتروپی مشترک[1] آن‌ها بدست می‌آید. سپس در صورتی که مقدار بدست آمده از حد آستانه‌ای[2] کمتر باشد، تغییر مفهوم تشخیص داده می‌شود. در این روش جزئیات رده‌بند و چگونگی بدست آوردن احتمال وقوع هر نمونه برای اضافه شدن به پنجره مشخص نشده است. همچنین این روش راهکاری برای تشخیص کلاس جدید ارائه نکرده است.

[1] Joint entropy
[2] Threshold

روش ارائه شده توسط مو و همکارانش [۴۰] از دو ماتریس خلاصه محلی و کلی برای تقریب داده‌ها استفاده می‌کند. ماتریس خلاصه کلی با استفاده از تمام مجموعه داده و برای تشخیص کلاس جدید تولید می‌شود، ولی ماتریس خلاصه محلی به ازای هر کدام از کلاس‌های موجود در داده‌های آموزشی ایجاد می‌شود و برای رده‌بندی داده‌ها به کار گرفته می‌شود. برای هر ماتریس بهینه‌ترین تقریب با استفاده از تجزیه مقادیر منفرد[۱] بدست می‌آید. هر یک از داده‌های آموزشی در ماتریس خلاصه کلی ضرب شده و محاسبه می‌شود. میانگین این اعداد بدست آمده به عنوان معیار شباهت در نظر گرفته می‌شود. هر یک از داده‌های آزمون را در این ماتریس ضرب کرده و با مقدار معیار شباهت بدست آمده مقایسه می‌کنیم. در صورتی که از مقدار معیار شباهت کمتر باشد به بافر اضافه شده و اگر بیشتر باشد رده‌بندی می‌شود. برای رده‌بندی، داده در ماتریس‌های خلاصه محلی مربوط به هر یک از کلاس‌ها ضرب شده و برای هر ماتریس که مقدار بیشتری بود، همان کلاس به عنوان برچسب داده در نظر گرفته می‌شود. اگر مقدار محاسبه شده از مقدار معیار شباهت کمتر باشد، ابتدا به بافر اضافه شده و بافر مرتب می‌گردد. سپس با استفاده از انحراف معیار نقطه جدایی دو طرف بافر بدست می‌آید و حداقل انحراف معیار به عنوان نقطه جدایی انتخاب می‌شود. حال در بخشی که انحراف معیار بیشتری دارد، تغییر مفهوم تشخیص داده شده و در بخشی که انحراف معیار کمتری دارد، کلاس جدید شناسایی می‌گردد. سپس با استفاده از داده‌های کلاس جدید ماتریس کلی به روز می‌شود. یکی از مشکلات این روش آن است که به تعدادی از داده‌هایی که شباهت کمتری با مقدار میانگین دارند، برچسب کلاس جدید زده و به درستی رده‌بندی نمی‌کند. همچنین در این روش در مورد تغییر مفهوم تصمیمی اتخاذ نشده است.

[۱] Singular value decomposition (SVD)

۲-۵ جمع‌بندی

در این فصل مفاهیمی همچون داده‌کاوی، رده‌بندی، جریان داده و رده‌بندی جریان داده معرفی شد. سپس به بررسی مدل‌های یادگیری نیمه‌نظارتی، روش‌های رده‌بندی جریان داده، مدل‌های پنجره‌گذاری و چند نرم‌افزار کار با جریان داده پرداخته شد. در نهایت نیز نمونه‌هایی از روش‌های رده‌بندی جریان داده، تشخیص تغییر مفهوم و تشخیص کلاس جدید در پژوهش‌های مرتبط با پایان نامه بیان شده و در مورد معایب آن‌ها توضیحاتی داده شد.

فصل ۳ : معرفی روش پیشنهادی

در جریان داده ممکن است با گذر زمان توزیع ویژگی یا برچسب نمونه‌ها تغییر کند. همچنین این امکان وجود دارد که یک کلاس جدید از داده‌ها تشکیل گردد. با تشخیص این موارد و به روز رسانی مدل‌های یادگیری می‌توان تا حد زیادی به بهبود دقت رده‌بندی کمک کرد. به همین دلیل در این فصل روشی ارائه خواهد شد که بتواند تغییر و تکامل مفهوم را تشخیص داده و با به روز رسانی مدل سعی در رده‌بندی درست داده‌ها داشته باشد. باید نمونه‌های کلاس جدید به سرعت شناسایی شده و داده‌هایی که به نظر می‌رسد متعلق به کلاس‌های موجود است، به رده‌بند داده شوند تا کلاس مربوطه را پیش‌بینی کند. زمانی که تعداد نمونه‌های کاندیدای کلاس جدید به حد مشخصی رسید یا اینکه تغییرات مفاهیم از حد آستانه‌ای بیشتر شد، باید مدل ارائه شده به روز رسانی شود. نتیجه کار باید به طور پیوسته و در طول جریان داده از دقت بالایی برخوردار باشد. فرض بر این است که برچسب کلاس برای تمام داده‌ها در دسترس نبوده و فقط مجموعه کوچکی از داده‌های برچسب‌دار اولیه وجود دارد.

همانطور که عنوان شد، در جریان داده فرض وجود برچسب برای همه داده‌ها یک فرض غیر واقع‌بینانه بوده و تعداد داده‌های برچسب‌دار نسبت به کل داده‌ها اندک است [۱۱]. بنابراین دسته نیمه‌نظارتی الگوریتم‌های یادگیری بیش از دسته نظارتی مورد نیاز است. به علاوه در بسیاری از برنامه‌های کاربردی عملیات برچسب‌گذاری هزینه‌بر و زمان‌بر بوده و برای افزایش کارایی و بهبود دقت می‌توان از داده‌های بدون برچسب استفاده کرد. به طور کلی در یادگیری نیمه‌نظارتی کلاس نمونه‌های بدون برچسب با توجه به نمونه‌های برچسب‌دار تعیین شده و این کار موجب ایجاد مطلوب‌تر رده‌بند می‌گردد. مدل ارائه شده از یک روش یادگیری نیمه‌نظارتی برای رده‌بندی جریان داده استفاده کرده که این روش علاوه بر تشخیص تغییر مفهوم، کلاس‌های جدید را نیز شناسایی کرده و به بهبود دقت رده‌بندی کمک می‌کند. روش ارائه شده می‌تواند به چهار زیر مسئله تجزیه شود:

- تشخیص کلاس جدید
- تغییر مفهوم

- رده‌بندی داده‌ها
- به روز رسانی مدل‌ها

برای تمام نمونه‌ها در جریان داده، یک فیلتر روی داده‌ها اعمال شده و تعلق داده به کلاس‌های موجود مشخص می‌شود. در صورتی که داده به یکی از کلاس‌های موجود تعلق داشته باشد، نمونه به رده‌بند تحویل داده می‌شود تا کلاس نمونه را پیش‌بینی نماید. در غیر این صورت نمونه مورد نظر به عنوان کاندیدای یک کلاس جدید تشخیص داده شده و در یک بافر قرار می‌گیرد. این بافر شامل کاندیداهایی است که به هیچ کلاسی اختصاص پیدا نکرده‌اند. حال با ورود نمونه به این بافر با استفاده از معیار q-NSC که در مورد آن توضیحاتی مطرح خواهد شد، کلاس جدید تشخیص داده شده و مدل به روز رسانی می‌شود. در صورتی که کلاس جدید تشخیص داده نشود، داده به پنجره‌ای برای تشخیص تغییر اضافه می‌گردد. با اضافه شدن داده به پنجره و محاسبه آنتروپی در صورتی که تغییرات آنتروپی پنجره از حد مشخصی بیشتر شد، تغییر مفهوم شناسایی شده و مدل به روز رسانی می‌شود. به روز رسانی مدل هم به این صورت است که مدل جدید جایگزین قدیمی‌ترین مدل موجود می‌گردد.

۳-۱ شمای کلی الگوریتم پیشنهادی

در این پژوهش از یک رده‌بند گروهی نیمه‌نظارتی شامل مدل‌های K-NN استفاده کرده‌ایم. گروهی بودن رده‌بند موجب می‌گردد بتوان علاوه بر داده‌های اولیه، از داده‌های جریان نیز برای رده‌بندی بهره برد و به مرور از داده‌های جدیدتر برای مدل‌سازی استفاده کرد. بدین ترتیب با توجه به این که مدل پیوسته با داده‌های جدید به روز رسانی می‌شود، دقت رده‌بندی بالا می‌رود. این روش نیازی به برچسب درست برای تمام داده‌ها نداشته و فقط به مجموعه کوچکی از داده‌های برچسب‌دار اولیه نیاز دارد. همچنین برای به روز رسانی رده‌بند، هیچ برچسب نمونه‌ای درخواست نمی‌کند. بنابراین روش ارائه شده برای مجموعه داده‌هایی که تعداد کمی از نمونه‌های آن برچسب درست در اختیار دارند، مناسب است.

روش پیشنهادی برای تشخیص کلاس جدید، کافی بودن تعداد نمونه‌های نزدیک به هم در بافر را بررسی می‌کند. این کار با محاسبه ضریب کلی q-NSC[1] انجام می‌شود [۱]. در صورتی که این ضریب یک مقدار مثبت شود؛ نشان دهنده آن است که نمونه‌ای که به بافر اضافه شده، به نمونه‌های درون بافر نزدیک‌تر و از نمونه‌های موجود کلاس‌ها دورتر است. بنابراین کلاس جدید تشخیص داده شده، باید مدلی با استفاده از این داده‌ها ساخته شود و مدل گروهی به روز رسانی گردد. روش پیشنهادی همچنین برای تشخیص تغییر از روشی مبتنی بر آنتروپی استفاده می‌کند. این تشخیص تغییر موجب می‌گردد درصورتی که در توزیع داده‌ها تغییر قابل توجهی اتفاق بیفتد مدل با به روز رسانی، خود را با این تغییر تطبیق داده و در بهبود دقت رده‌بندی موثر عمل کند.

در ادامه الگوریتم پیشنهادی به صورت گام به گام توضیح داده خواهد شد. در ابتدا برای مطالعه ساده‌تر و همچنین روشن بودن تمامی زوایای الگوریتم، کلیه پارامترهای به کار رفته در جدول ۳-۱ گردآوری شده است.

جدول ۳-۱.پارامترهای موجود در الگوریتم پیشنهادی

پارامترهای الگوریتم	مفهوم
K	تعداد خوشه‌ها
CS	تعداد قطعه داده
M	تعداد مدل‌های رده‌بند گروهی
Window size	اندازه پنجره
Buffer size	اندازه بافر
q’	حداقل تعداد داده‌های بافر برای بررسی کلاس جدید

[1] q-Neighborhood Silhouette Coefficient

پس از مشخص شدن پارامترهای به کار گرفته شده در الگوریتم مذکور، به بررسی شبه کد مربوطه پرداخته می‌شود. مطابق با شکل ۳-۱ در ابتدا رده‌بند گروهی شامل مدل‌هایی است که روی داده‌های آموزشی اولیه آموزش دیده‌اند. سپس با ورود داده‌های جریان ابتدا بررسی می‌گردد داده دور افتاده است یا خیر. به ازای هر کدام از مدل‌ها دور افتاده بودن داده بررسی می‌شود. درصورتی که در یکی از مدل‌ها داده درون خوشه‌ای قرار گیرد، آن داده دور افتاده در نظر گرفته نشده و برای رده‌بندی باید روی برچسب داده‌های خوشه‌ای که داده جریان در آن قرار گرفته، رای اکثریت انجام گیرد. با انجام رای اکثریت در واقع محتمل‌ترین برچسب به داده تعلق می‌گیرد. البته درصورتی که داده در بیش از یک مدل درون خوشه قرار گیرد، باید علاوه بر انجام رای اکثریت بر روی خوشه مذکور، یک بار هم بر روی تمام مدل‌هایی که داده درون خوشه واقع شده رای اکثریت صورت پذیرد. در صورتی که داده در هیچ‌یک از مدل‌ها درون خوشه‌ای قرار نگیرد؛ با در نظر گرفتن یک بافر، هر کدام از داده‌ها به این بافر اضافه شده و درصورتی که تعداد داده‌های موجود در بافر به تعداد مشخصی رسید، با محاسبه ضریب q-NSC روی داده‌های بافر، تشخیص کلاس جدید صورت می‌گیرد. درصورتی که کلاس جدید شناسایی شود، مدل گروهی به روز رسانی شده و داده‌های بافر حذف می‌گردد. در غیر این صورت داده ورودی به بخش تشخیص تغییر مفهوم منتقل می‌شود. در این بخش با در نظر گرفتن دو پنجره، احتمال تعلق هر کدام از این داده‌ها به ازای هر کدام از مدل‌ها محاسبه شده و میانگین آن‌ها به دو پنجره وارد می‌شوند. درصورتی که دو پنجره پر شده باشد، اختلاف آنتروپی دو پنجره بدست آمده و اگر اختلاف آن‌ها از حد آستانه‌ای بیشتر باشد، تغییر مفهوم شناسایی می‌گردد. زمانی که تغییر تشخیص داده شود، مدل گروهی به روز رسانی شده و داده‌های دو پنجره حذف می‌گردد. در شکل ۳-۲ شمای کلی الگوریتم پیشنهادی نشان داده شده است.

شبه کد الگوریتم پیشنهادی

Input: $data\ stream, K, M, q', CS$

Data stream of train instances partitioned into CS chunks

For *each chunk* CS :

M_i ⟵ Create Classifier C_i (K-means++)

For *each model* M_i :

Discarded data points after saving pseudopoints in each cluster

(centers, radius, number of data of each class)

For *each test instance* t :

if t *is inside the decision boundary of model* M_i

C(t) ⟵ Classify(t)

else

add t to buffer

result ⟵ calculate q-NSC(t)

if result is positive for at least q' data

A new class detected and update classifier

Remove all data from buffer

else

add t to window

if *window is full*

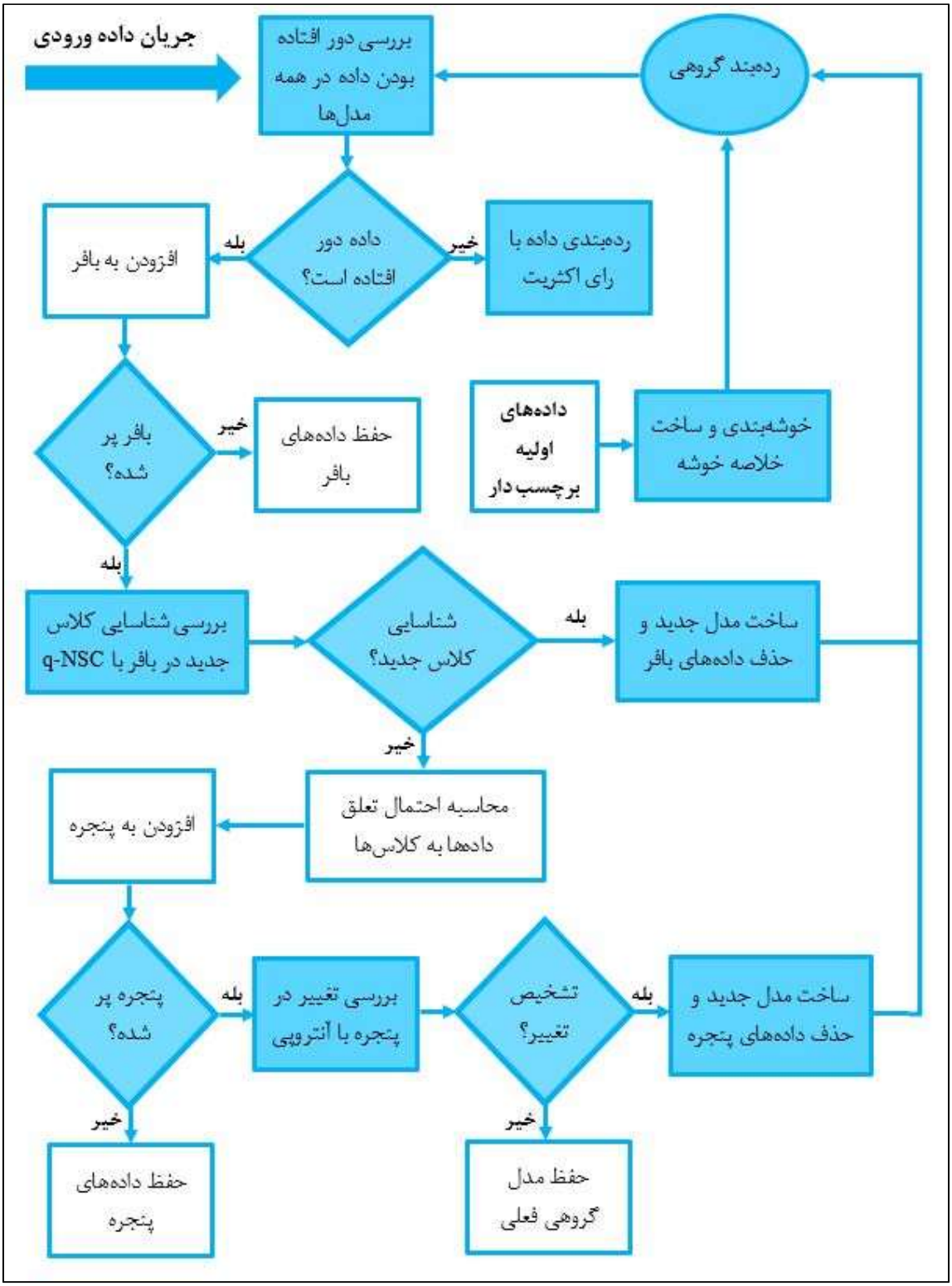
Change detection is checked by Entropy based ensemble

if change *detected*

Update classifier and remove all data from window

Output: class label for each data

شکل ۳-۱.شبه کد الگوریتم پیشنهادی



شکل ۳-۲.شمای کلی روش پیشنهادی

۳-۲ آموزش و رده‌بندی

مجموعه داده‌ای که برای رده‌بندی به الگوریتم داده می‌شود، شامل دو بخش است: 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 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[۱] می‌گوییم، با استفاده از یک الگوریتم خوشه‌بندی مانند K-means یا DBSCAN [۱۰] 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 کلاس‌ها است. شعاع یک خوشه، فاصله بین مرکز و دورترین نقطه داده درون خوشه در نظر گرفته می‌شود. معیار فاصله در نظر گرفته شده، فاصله اقلیدسی[۲] است.

در این پژوهش 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 تعدادی قطعه داده تقسیم شده و روی هر قطعه داده i با استفاده از الگوریتم K-means++ خوشه‌بندی انجام می‌شود و مدل m_i بدست می‌آید. مقدار k برای تمام قطعه داده‌ها یکسان در نظر گرفته شده است. پس از حذف داده‌ها، هر مدل m_i شامل مجموعه‌ای از k خلاصه داده است.

---

۱ Chunk

۲ Euclidean distance

پس از حذف داده‌ها با ورود هر داده از جریان داده ورودی و با توجه به ویژگی‌های داده و شعاع خوشه، بررسی می‌گردد که داده درون خوشه قرار می‌گیرد یا خیر. با فرض داشتن تنها یک رده‌بند، در صورتی که داده درون یکی از خوشه‌ها قرار گیرد، رای اکثریت روی خوشه گرفته شده و کلاس با بیشترین تعداد تکرار در خوشه مذکور، به عنوان کلاس داده جریان در آن مدل تلقی می‌گردد. ولی با توجه به گروهی بودن روش رده‌بندی، دوباره بر روی مدل‌هایی که داده را درون خوشه تشخیص داده‌اند، رای اکثریت گرفته شده و کلاس با بیشترین تکرار در مدل‌های موجود، به عنوان کلاس داده جریان در نظر گرفته می‌شود. درصورتی که داده برای هیچ یک از مدل‌ها درون خوشه‌ای قرار نگیرد، با محاسبه ضریب q-NSC تشخیص کلاس جدید برای نمونه بررسی شده و درصورتی که کلاس جدید شناسایی شد، مدل گروهی به روز رسانی می‌شود. در غیر این صورت، داده به بخش تشخیص تغییر مفهوم منتقل می‌شود. در این سیستم با استفاده از روشی مبتنی بر آنتروپی تشخیص تغییر برای نمونه بررسی شده و درصورتی که تغییر تشخیص داده شد، مدل گروهی به روز رسانی می‌شود.

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 اینکه مقدار k برای اجرای الگوریتم K-NN یک در نظر گرفته شود، کلاسی که برای x در نظر گرفته می‌شود، بیشترین تعداد تکرار کلاس در خلاصه داده h می‌باشد. رده‌بندی داده x با استفاده از رده‌بند گروهی M و با گرفتن رای اکثریت از بین همه رده‌بندها انجام می‌شود. در صورتی که مقدار k غیر یک باشد، باید از بین k خلاصه داده رای اکثریت 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 می‌گردد.

۳-۳ تشخیص کلاس جدید

هر داده جریان ابتدا توسط رده‌بند گروهی M بررسی می‌گردد. اگر نمونه داخل مرز تصمیم باشد، طبق بخش ۳-۲ رده‌بندی می‌گردد. در غیر این صورت، به عنوان یک داده دور افتاده در نظر گرفته می‌شود. اساس تشخیص کلاس جدید این است که یک داده به داده‌های کلاس خودش نزدیکتر بوده (همبستگی[۱] داشته باشد) و از داده‌های سایر کلاس‌ها دورتر باشد (جدا[۲] باشد). بنابراین اگر یک کلاس جدید در جریان داده وجود داشته باشد، باید نمونه‌های متعلق به آن کلاس از نمونه‌های کلاس‌های موجود دور بوده و به نمونه‌های کلاس جدید نزدیک باشند. به دلیل اینکه داده‌های دور افتاده خارج از مرز تصمیم واقع شده‌اند، از کلاس‌های موجود دور هستند. بنابراین ویژگی جدایی برای یک کلاس جدید با استفاده از این داده‌های دور افتاده برآورده شده و این نمونه‌ها به صورت موقت در یک بافر ذخیره می‌شوند تا ویژگی همبستگی آن‌ها نیز بررسی شود. با ورود هر داده جریان، همبستگی داده‌های بافر مورد بررسی قرار می‌گیرد تا مشخص شود تعداد داده‌های دور افتاده‌ای که به یکدیگر نزدیک هستند کافی است یا خیر. این کار با محاسبه ضریب کلی q-NSC انجام می‌شود [۴۳]. این ضریب بر پایه q و همسایگی c بوده و مجموعه‌ای از q نمونه نزدیکتر به نمونه x که دارای برچسب کلاس c هستند، تعریف می‌شود.

در (رابطه ۳-۱) نحوه محاسبه ضریب q-NSC آمده است که در این رابطه $\overline{D}_{c_{out,q}}(x)$ میانگین فاصله داده دور افتاده x تا q نزدیکترین همسایه‌اش است. $\overline{D}_{c_{min,q}}(x)$ حداقل فاصله داده x تا همه کلاس‌های موجود یا به عبارت دیگر نزدیکترین کلاس به داده x است.

$$q - NSC(x) = \frac{\overline{D}_{c_{min,q}}(x) - \overline{D}_{c_{out,q}}(x)}{max(\overline{D}_{c_{min,q}}(x), \overline{D}_{c_{out,q}}(x))} \quad \text{(رابطه ۳-۱)}$$

[۱] Cohesion

[۲] Separation

این رابطه به صورت همزمان همبستگی و جدایی را می‌سنجد. مقدار حاصل عددی بین ۱- و۱+ است که مقدار مثبت‌تر به معنی نزدیکتر بودن داده‌های دور افتاده (همبستگی) بوده و از کلاس‌های موجود دورتر (جدایی) است. باید حداقل q' تا از داده‌های دور افتاده موجود باشد تا یک کلاس جدید تشخیص داده شود.

۳-۴ تشخیص تغییر

آنتروپی در نظریه اطلاعات معیاری عددی از میزان تصادفی بودن یک متغیر تصادفی است که به صورت رابطه (۳-۲) تعریف شده است [۱۵]. 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. 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 شده و N داده بعدی وارد پنجره دوم می‌شود. با پر شدن پنجره‌ها، داده‌های بعدی یک به یک به پنجره B_2 وارد شده و از طرف مقابل قدیمی‌ترین داده از پنجره B_1 خارج 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 (۳-۳) استفاده می‌کنیم.

$$H(x) = -\sum_{i=1}^{n} P_i \, log_2 \, P_i \qquad (۳-۲)$$

$$H(x) = -\sum_{i=1}^{c} \sum_{x} P_i^x \, log_2 \, P_i^x \qquad (۳-۳)$$

که در این روابط x متغیر تصادفی گسسته، P_i احتمال رخداد x_i ، $P(x)$ تابع احتمال چگالی x و c تعداد کلاس‌ها 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 بدست آمده برای هر مدل، میانگین گرفته و مقدار P_i^x محاسبه می‌گردد. این مدل ارائه شده، می‌تواند دو نوع تغییر مفهوم ناگهانی و تدریجی را تشخیص دهد.

۳-۵ به روز رسانی مدل گروهی رده‌بند

مدل‌های رده‌بند باید با تشخیص کلاس جدید و یا شناسایی تغییر مفهوم به روز رسانی شوند. زمانی که کلاس جدید تشخیص داده می‌شود، رده‌بند با استفاده از داده‌های بافر به روز می‌شود. این به روز رسانی به این صورت است که یک مدل جدید با استفاده از داده‌های بافر ساخته شده و جایگزین قدیمی‌ترین مدل از میان مدل‌های موجود می‌گردد. با به روز رسانی مدل داده‌های بافر حذف می‌شود.

زمانی که یک تغییر مفهوم شناسایی می‌گردد، رده‌بند با استفاده از داده‌های دو پنجره به روز می‌شود. به این صورت که یک مدل جدید با استفاده از داده‌های دو پنجره ساخته شده و جایگزین قدیمی‌ترین مدل می‌شود. این کار تضمین می‌کند دقیقا تعداد مشخصی مدل در هر زمان وجود داشته باشد. پس از به روز رسانی مدل، داده‌های دو پنجره حذف می‌شود.

به این ترتیب روش پیشنهادی از داده‌های جریان نیز برای آموزش رده‌بند استفاده می‌کند و مساله تغییر مفهوم هم با به روز رسانی مدل گروهی و با استفاده از مفاهیم اخیر پاسخ داده می‌شود. زمانی که توزیع داده‌های یک کلاس تغییر کند، تغییر مفهوم رخ می‌دهد. بنابراین لازم است با به روز رسانی رده‌بند، داده‌هایی که قبلا به عنوان داده دور افتاده شناخته می‌شده، اکنون به عنوان نمونه‌ای از یک کلاس تلقی کنیم. این شناسایی تغییرات در بهبود عملکرد رده‌بندی موثر است. درصورتی که تغییری در توزیع داده‌ها اتفاق نیفتد، رده‌بند قادر است با مدل‌های موجود رده‌بندی را انجام دهد. به این دلیل به روز رسانی رده‌بند تنها زمانی که یک تغییر مفهوم شناسایی شود، انجام می‌شود. همچنین با تشخیص کلاس جدید لازم است رده‌بند، داده‌هایی که تاکنون آن‌ها را دور افتاده فرض می‌کرده، به عنوان یک

کلاس جدید تلقی کند. بنابراین تنها زمانی مسئله به روز رسانی رده‌بند مطرح می‌شود که یا کلاس جدیدی شناسایی شود و یا اینکه در داده‌ها تغییر محسوسی تشخیص داده شود.

۳-۶ جمع‌بندی

در این فصل روش پیشنهادی معرفی و تشریح گردید. همانطور که اشاره شد در این روش به بررسی تشخیص کلاس جدید و تغییر مفهوم پرداخته شد. برای تشخیص کلاس جدید از یک رابطه که همزمان جدایی و همبستگی داده‌ها را می‌سنجد، استفاده شد. همچنین از روشی مبتنی بر آنتروپی برای تشخیص تغییر مفهوم بهره برده شد. پارامترهای استفاده شده در الگوریتم‌های این بخش در قالب جدولی برای روشن‌تر شدن مفهوم الگوریتم ارائه گردید. همچنین در این بخش شمای کلی از الگوریتم به منظور درک بهتر موضوع رسم شد که به صورت دقیق مرحله به مرحله فازهای پیاده‌سازی را بیان نموده است. در روش پیشنهادی به دلیل کارایی بیشتر در زمینه جریان داده، از یادگیری نیمه‌نظارتی به جای روش نظارتی استفاده شده است. همچنین گروهی بودن رده‌بند موجب می‌گردد بتوان از داده‌های جریان برای رده‌بندی بهره برد. روش ارائه شده از تغییر آنتروپی برای تشخیص تغییر استفاده کرده و می‌تواند دو نوع تغییر مفهوم ناگهانی و تدریجی را در جریان داده تشخیص دهد. همچنین با ارائه این روش، کلاس جدید و تغییر مفهوم به سرعت شناسایی شده و با به روز رسانی مدل گروهی، دقت رده‌بندی را افزایش می‌دهد.

فصل ۴ : پیاده‌سازی و ارزیابی روش پیشنهادی

۴-۱ تنظیمات و راه‌اندازی سیستم

در این فصل به بیان پیاده‌سازی و ارزیابی راهکار پیشنهادی و جزییات مربوط به آن پرداخته می‌شود. در ابتدا به معرفی مجموعه داده‌های به کار گرفته شده پرداخته و سپس به بیان جزئیات مربوط به فرآیند پیاده‌سازی اقدام می‌کنیم. در ادامه مراحل راه‌اندازی سیستم و تنظیمات مورد نیاز برای اجرای الگوریتم رده‌بندی جریان داده تکاملی را شرح می‌دهیم. انجام آزمایش‌ها و ارزیابی نتایج و در نهایت مقایسه‌ی راهکار پیشنهادی با تعدادی از روش‌های مشابه نیز در انتهای این فصل صورت می‌پذیرد. لازم به ذکر است که تمامی مراحل پیاده‌سازی (ساخت مجموعه ویژگی‌ها و سیستم پیشنهادی) به زبان برنامه‌نویسی جاوا و در محیط NetBeans IDE انجام پذیرفت.

۴-۲ مجموعه داده

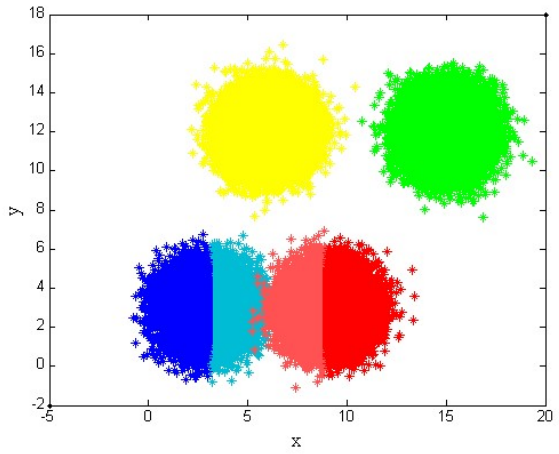
به منظور ارزیابی روش پیشنهادی از مجموعه داده‌های مختلفی استفاده شده است. ۳ مجموعه داده ابتدایی جدول زیر از وب‌سایت UCI [۱] گرفته شده و دو مجموعه داده دیگر ساختگی است.

جدول ۴-۱. مشخصات مجموعه داده‌های استفاده شده برای ارزیابی عملکرد روش پیشنهادی

نام مجموعه داده	تعداد نمونه	تعداد کلاس	تعداد ویژگی
Forest Cover	۱۰۰,۰۰۰	۷	۵۴
Shuttle	۴۰,۰۰۰	۷	۹
Skin Segmentation	۱۰۰,۰۰۰	۲	۳
Gaussian Generator	۱۵۰,۰۰۰	۴	۲
Random RBF Generator Drift	۱۰۰,۰۰۰	۵	۱۷

[۱] https://archive.ics.uci.edu/ml/datasets.php.

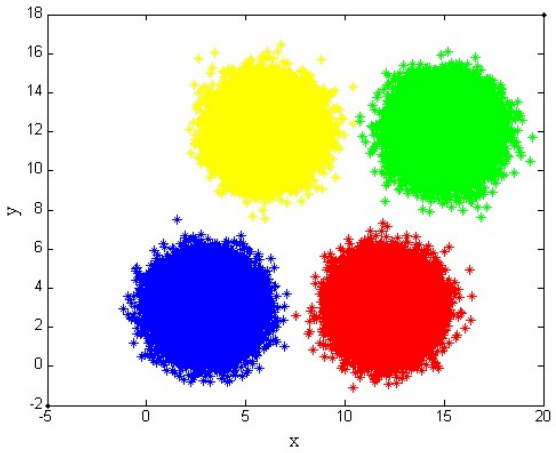
در جدول ۴-۱ مشخصات مجموعه داده‌های استفاده شده برای ارزیابی روش پیشنهادی با سایر روش‌های رده‌بندی آورده شده است. مجموعه داده Gaussian Generator با استفاده از توزیع گوسی با ۴ کلاس و ۲ ویژگی ساخته شده است. این مجموعه داده با ثابت بودن تعداد کلاس و ویژگی‌ها در سه مدل ساخته شده است. در مدل اول فقط تغییر مفهوم در داده‌ها بوجود آمده و کلاس جدید در نظر گرفته نشده و مطابق شکل ۴-۱ تعدادی از نمونه‌های دو کلاس آبی و قرمز همپوشانی دارند. برای ایجاد تغییر مفهوم، داده‌های سمت چپ کلاس آبی و سمت راست کلاس قرمز به عنوان داده‌های آموزشی و برای داده‌های جریان، داده‌های سمت راست کلاس آبی و سمت چپ کلاس قرمز به الگوریتم داده شده است. برای دو کلاس سبز و زرد رنگ هم بخشی از نمونه‌های این دو کلاس به صورت تصادفی به عنوان داده آموزشی در نظر گرفته شده و بقیه داده‌های آن‌ها به عنوان داده جریان به الگوریتم داده شده است. هدف استفاده از این مجموعه داده، بررسی عملکرد الگوریتم با وجود تغییر مفهوم و همپوشانی دو کلاس به صورت همزمان است.



شکل ۴-۱.نمایش مجموعه داده ساختگی با استفاده از توزیع گوسی

در مدل دوم فقط کلاس جدید در داده‌ها در نظر گرفته شده است. مطابق شکل ۴-۲ بخشی از داده‌های سه کلاس آبی، قرمز و زرد به عنوان داده‌های آموزشی به الگوریتم داده شده و برای بررسی عملکرد

روش پیشنهادی در زمان وقوع کلاس جدید، داده‌ای از کلاس سبز برای آموزش به الگوریتم داده نشده است.



شکل ۴-۲.نمایش مجموعه داده ساختگی با استفاده از توزیع گوسی و بدون تغییر مفهوم

در مدل سوم همانند مدل اول تغییر مفهوم در داده‌ها وجود دارد ولی کلاس جدید نیز در داده‌ها در نظر گرفته شده است. مطابق شکل ۴-۱ و آنچه در مدل اول بیان شد، برای سه کلاس آبی، قرمز و زرد داده‌ها به همان ترتیب به الگوریتم داده شده است. تنها تغییر در داده‌های کلاس سبز است که داده‌ای از کلاس سبز برای آموزش به الگوریتم داده نشده است.

مجموعه داده Forest Cover شامل ۵۴ ویژگی و ۷ کلاس بوده که ما برای ارزیابی نتایج، پس از نرمال‌سازی از ۱۰۰ هزار نمونه آن استفاده کرده‌ایم. مجموعه داده Shuttle شامل ۹ ویژگی و ۷ کلاس است که ۴۰ هزار نمونه برای ارزیابی نتایج انتخاب شده است. در مجموعه داده Skin Segmentation از ۱۰۰ هزار نمونه به عنوان داده آموزشی و داده جریانی استفاده شده است. این سه مجموعه داده از وب‌سایت UCI گرفته شده و مجموعه داده آخر با استفاده از Random RBF Generator Drift در MOA با ۵ کلاس و ۱۷ ویژگی ساخته شده است. MOA یک محیط نرم‌افزاری برای اجرای الگوریتم‌ها و اجرای آزمایشات برای یادگیری برخط جریان داده است [۱۶]. مجموعه داده Random RBF Generator به این صورت ساخته می‌شود که ابتدا تعدادی مرکز تصادفی تولید می‌شود. به‌طوری‌که هر

مرکز دارای موقعیت، انحراف معیار، برچسب کلاس و وزن می‌باشد. نمونه‌های جدید با انتخاب تصادفی یک مرکز تولید شده و با توجه به وزن هر یک از مراکز، احتمال انتخاب مراکز، متفاوت خواهند بود. مرکز انتخاب شده همچنین برچسب کلاس نمونه را تعیین می‌کند. تغییر مفهوم با تغییر در مراکز و با یک سرعت ثابت در مجموعه داده ایجاد می‌شود [۱۴]. 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 ۶۰۰۰ نمونه به عنوان داده آموزشی برای ساخت مدل‌های اولیه در نظر گرفته شده و بقیه نمونه‌ها به عنوان داده‌های جریان به الگوریتم داده شده است.

۴-۳ پیاده‌سازی روش ارائه شده

به طور کلی سیستم پس از دریافت مجموعه داده مورد نظر، فرآیند رده‌بندی جریان داده را آغاز می‌کند. ابتدا پارامترهای الگوریتم مشخص شده، سپس داده‌ها خوشه‌بندی و خلاصه‌سازی می‌شوند. پس از خلاصه‌سازی داده‌ها، هر داده جریان مطابق با الگوریتمی که در فصل سوم ارائه شد، رده‌بندی می‌گردد.

در جدول ۴-۲ خلاصه‌ای از مقادیر پارامترهای مورد نیاز برای راه‌اندازی الگوریتم رده‌بندی مورد نظر نشان داده شده است. برای ساخت هر مدل، یک قطعه داده معادل با ۱۰۰۰ داده آموزشی اولیه را خوشه‌بندی می‌کنیم. در مجموع از ۶۰۰۰ داده برای آموزش مدل رده‌بند استفاده می‌شود. تعداد خوشه‌ها برای خوشه‌بندی داده‌ها متناسب با اندازه جریان داده بین ۴ تا ۱۰ خوشه تعیین می‌شود. پس از خوشه‌بندی و خلاصه‌سازی داده‌ها، فرآیند رده‌بندی داده‌های جریان آغاز می‌گردد. همچنین حداقل تعداد داده‌های همسایه برای بررسی کلاس جدید با استفاده از ضریب q-NSC ۵۰ داده در نظر گرفته

شده است. لازم به ذکر است که در انتخاب مقادیر پارامترهای اندازه پنجره و اندازه بافر، متناسب با اندازه مجموعه داده مورد استفاده عمل می‌کنیم.

جدول ۴-۲.مقادیر پارامترها در پیاده‌سازی

پارامتر	توضیح	مقدار
K	تعداد خوشه‌ها	۴ تا ۱۰
CS	تعداد قطعه داده	۶
Chunk size	اندازه قطعه داده	۱۰۰۰
M	تعداد مدل‌های رده‌بند گروهی	۶
Window size	اندازه پنجره	یک دهم اندازه جریان داده
Buffer size	اندازه بافر	یک دهم اندازه جریان داده
q’	حداقل تعداد همسایه برای بررسی کلاس جدید	۵۰

۴-۳ آزمایش‌ها و ارزیابی نتایج

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

$$Accu(y,\hat{y}) = \frac{1}{n_{samples}} \sum_{i=0}^{n_{samples}-1} 1(\hat{y_i} = y_i)$$ (رابطه ۴-۱)

$$PR = \frac{TP}{TP+FP}$$ (رابطه ۴-۲)

$$RE = \frac{TP}{TP+FN}$$ (رابطه ۴-۳)

$$F1 = \frac{2*PR*RE}{PR+}$$ (رابطه ۴-۴)

در (رابطه ۴-۱) منظور از $\hat{y_i}$ مقدار پیش‌بینی شده نمونه i ام و y_i مقدار واقعی نمونه i و n تعداد نمونه‌های موجود است. در (رابطه ۴-۲) و (رابطه ۴-۳) منظور از TP تعداد نمونه‌هایی است که به درستی توسط رده‌بند به کلاس C_i انتساب یافته و FP یعنی تعداد نمونه‌هایی که متعلق به سایر کلاس‌ها بوده و به اشتباه توسط رده‌بند به کلاس C_i منسوب شده‌اند. FN تعداد نمونه‌هایی است که متعلق به کلاس C_i بوده ولی توسط رده‌بند به سایر کلاس‌ها تعلق گرفته است. معیار TN تعداد نمونه‌هایی است که متعلق به کلاس C_i نبوده‌اند و توسط رده‌بند به این کلاس اختصاص داده نشده‌اند. در (رابطه ۴-۴) هم معیار F1 با استفاده از دو معیار دقت و یادآوری بدست می‌آید.

ابتدا مقایسه‌ای بین روش پیشنهادی با دو روش OBA و HAT 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 حداکثر تعداد تقسیم و حذف برخی گره‌های درخت در روش HT ، سعی در کوچکتر کردن درخت برای سازگاری سریعتر الگوریتم با تغییرات موجود در جریان داده کرده است [۱۷]. OBA (OzaBagAdwin) یک روش گروهی برای تشخیص تغییر به همراه تخمین وزن برای بهبود عملکرد است. تشخیص تغییر با مقایسه میانگین مقادیر نمونه‌هایی که قبل و بعد از اضافه شدن نمونه i ام به پنجره بدست می‌آیند، انجام می‌گیرد. زمانی که تغییر تشخیص داده می‌شود، رده‌بند جدید جایگزین بدترین رده‌بند در مدل گروهی می‌شود [۱۴].

جدول ۴-۳ نتایج بدست آمده از رده‌بندی روش پیشنهادی و دو روش OBA و HAT را بر روی مجموعه داده Forest Cover نشان می‌دهد. در این جدول میزان دقت، یادآوری و همچنین F1 هر کدام از این سه روش به تفکیک برای هر کلاس آمده است. در این مجموعه داده، تغییرات داده‌های کلاس ۴ به صورتی است که در یک پنجره برای بررسی آنتروپی قرار نمی‌گیرند و به همین دلیل موجب اختلاف آنتروپی زیاد نمی‌شوند. همین امر باعث شده تغییرات کلاس ۴ در مواردی به خوبی تشخیص داده نشده

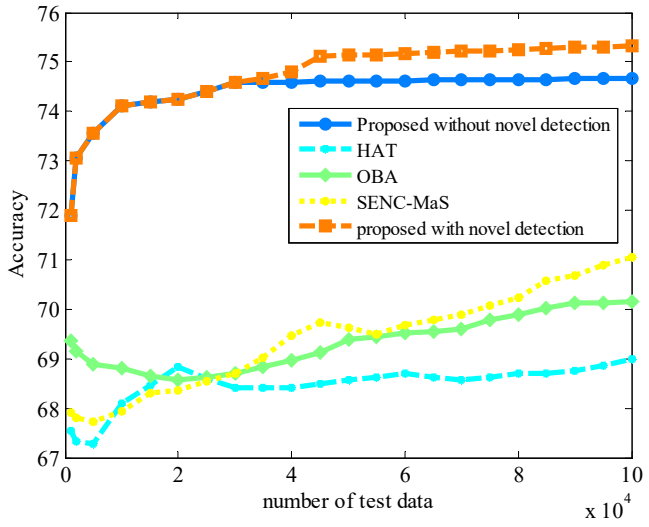
و کلاس دیگری به جای آن برچسب گرفته که باعث زیاد شدن FN برای این کلاس شده و معیار یادآوری را تا حد زیادی کاهش داده است. همانطور که مشاهده می‌شود، روش پیشنهادی برای بیشتر کلاس‌های موجود در مجموعه داده در هر سه معیار ارزیابی بالاتر بوده و در کل عملکرد بهتری داشته است.

جدول ۴-۳.مقایسه روش پیشنهادی و دو روش OBA و HAT بر روی Forest Cover

روش / معیار		روش پیشنهادی	OBA	HAT
PR	C1	**۹۹/۳**	۶۶/۴	۵۹/۸
	C2	۶۷/۱	**۶۸/۹۹**	۶۶
	C3	**۱۰۰**	۷۹/۶۵	۷۷/۹۸
	C4	**۱۰۰**	۴۸/۳۶	۴۷/۳۳
	C5	**۱۰۰**	۱۷/۷۶	۱۷/۵۳
	C6	**۱۰۰**	۳۰/۵۹	۲۸/۹۴
	C7	**۱۰۰**	۵۳/۹۸	۵۵/۶۶
RE	C1	۵۳/۵۲	۴۵/۹۱	**۵۷/۰۹**
	C2	**۱۰۰**	۸۹/۹۴	۷۹/۲
	C3	۳۵/۱۹	**۵۶/۶۹**	۵۲/۰۷
	C4	۵/۴۶	**۴۲/۲۱**	۳۹/۹۶
	C5	**۵۲/۱۵**	۹/۰۵	۱۴/۵۶
	C6	۳۶/۲۳	**۸۰/۶۴**	۷۴/۸
	C7	**۶۰/۶۱**	۴۷/۵۳	۵۹/۰۶
F1	C1	**۶۹/۵۵**	۵۴/۲۹	۵۸/۴۱
	C2	**۸۰/۳۱**	۷۸/۰۸	۷۲
	C3	۵۲/۰۶	**۶۶/۲۴**	۶۲/۴۴
	C4	۱۰/۳۵	**۴۵/۰۸**	۴۳/۳۳
	C5	**۶۸/۵۵**	۱۱/۹۹	۱۵/۹۱
	C6	**۵۳/۱۹**	۴۴/۳۵	۴۱/۷۳
	C7	**۷۵/۴۷**	۵۰/۵۵	۵۷/۳۱

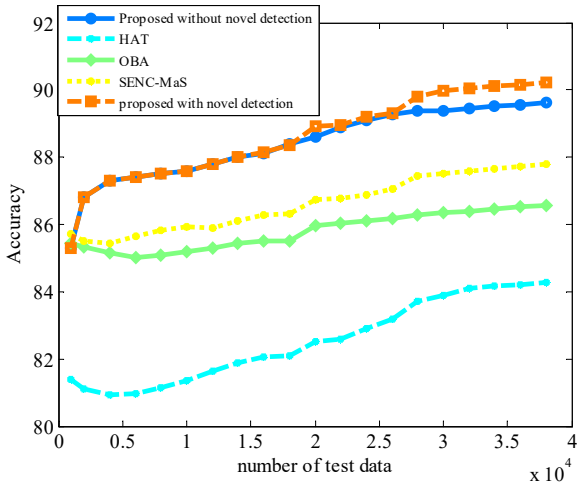
در ادامه روش پیشنهادی بر مجموعه داده‌های مختلف با ۳ روش OBA، HAT و SENC-MaS مقایسه می‌گردد. روش SENC-MaS [۴۰] به عنوان آخرین پژوهش در بخش بررسی پژوهش‌های انجام شده در فصل دو تشریح گردیده است. همانطور که اشاره شد، این روش از دو ماتریس خلاصه برای تقریب داده‌ها استفاده می‌کند و برای هر ماتریس با استفاده از تجزیه مقادیر منفرد بهینه‌ترین تقریب را بدست می‌آورد. همچنین از میانگین اعداد بدست آمده از این ماتریس به عنوان معیار شباهت استفاده کرده و برای تشخیص کلاس جدید بهره می‌برد. این روش به این دلیل که روش اخیری است که کلاس جدید را تشخیص می‌دهد، برای مقایسه با روش پیشنهادی انتخاب شده است.

در شکل ۴-۳ مقایسه این سه روش با روش پیشنهادی بر روی مجموعه داده Forest Cover با معیار ارزیابی صحت نشان داده شده است که در هر مرحله ۵۰۰۰ داده به مجموعه داده جریان اضافه شده است. برای اجرای الگوریتم‌ها بر روی این مجموعه داده، از داده ۱۸۰۰۰۰ تا داده ۲۸۰۰۰۰ مجموعه داده اصلی انتخاب شده است. در این مجموعه داده، داده‌های دارای برچسب کلاس ۴ از ابتدای مجموعه داده مذکور تا داده ۲۱۹۰۰۰ وجود نداشته و پس از آن پدیدار می‌گردد. در نتایج برای نمایش بهتر عملکرد روش پیشنهادی، به دو صورت در نمودار نشان داده شده است. نمودار آبی رنگ روش پیشنهادی با سیستم تشخیص تغییر مفهوم بوده و مکانیزم شناسایی کلاس جدید ندارد. نمودار نارنجی رنگ روش پیشنهادی با سیستم تشخیص تغییر و کلاس جدید به صورت توام می‌باشد. معیار صحت روش پیشنهادی در تمامی مراحل بیشتر بوده و عملکرد خوبی داشته است. همچنین همانطور که در شکل ملاحظه می‌شود، در داده جریان ۳۹۰۰۰ کلاس جدید شناسایی شده و به همین دلیل در نهایت معیار صحت روش پیشنهادی دارای تشخیص کلاس جدید از روش پیشنهادی دیگر اندکی بیشتر شده است.



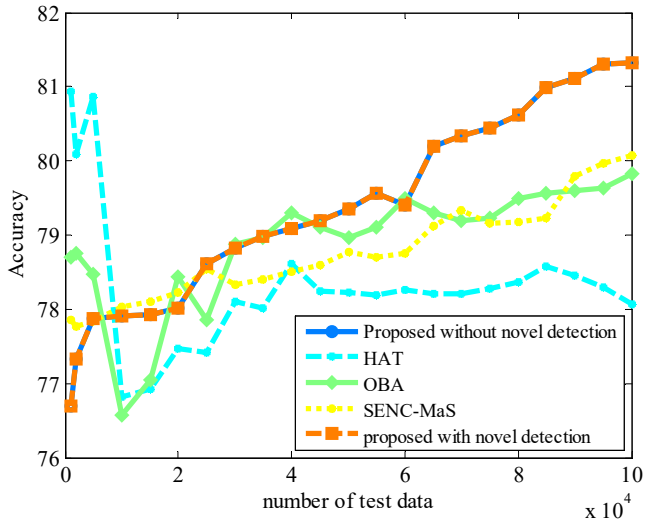
شکل ۴-۳.نمودار تغییر صحت برای مجموعه داده Forest Cover

در شکل ۴-۴ روش پیشنهادی با سه روش OBA، HAT و SENC-MaS بر روی مجموعه داده Shuttle مقایسه شده است که در هر مرحله ۲۰۰۰ داده به مجموعه داده جریان اضافه شده است. در این مجموعه داده، داده‌های دارای برچسب کلاس ۲ در مجموعه داده آموزشی وجود نداشته و از داده ۱۰۰۰ جریان داده به تدریج پدیدار می‌گردد. به طوری که تعداد داده‌های دارای برچسب کلاس ۲ در داده جریان حدود ۱۸۰۰۰ به تعداد قابل توجهی رسیده و کلاس جدید توسط روش پیشنهادی تشخیص داده می‌شود. همچنین در این مجموعه داده، داده‌های دارای برچسب کلاس ۷ در مجموعه داده نبوده و در داده جریان حدود ۲۵۰۰۰ پدیدار می‌گردد، که روش پیشنهادی در داده ۲۷۰۰۰ این کلاس را تشخیص داده است. صحت روش پیشنهادی در تعداد داده جریان کمتر از ۲۰۰۰ با روش OBA و SENC-MaS برابر است، ولی به تدریج با زیاد شدن داده‌های جریان خیلی سریع صحت روش پیشنهادی بالا رفته و از آن دو روش فاصله می‌گیرد. همانطور که مشاهده می‌شود، تشخیص کلاس جدید در داده‌ها در دو مقطع باعث افزایش صحت روش پیشنهادی دارای تشخیص کلاس جدید نسبت به روش پیشنهادی دیگر شده است.



شکل ۴-۴.نمودار تغییر صحت برای مجموعه داده Shuttle

در شکل ۴-۵ روش پیشنهادی با سه روش OBA، HAT و SENC-MaS بر روی مجموعه داده Random RBF Generator Drift مقایسه شده است که در هر مرحله ۵۰۰۰ داده به مجموعه داده جریان اضافه شده است. همانطور که مشاهده می‌شود، صحت روش پیشنهادی در تعداد داده جریان کمتر از ۶۰۰۰۰ با روش OBA برابر بوده و به دو روش SENC-MaS و HAT نزدیک است، ولی به تدریج با زیاد شدن داده‌های جریان خیلی سریع صحت روش پیشنهادی بالا رفته و از سه روش مذکور فاصله می‌گیرد. همانطور که در این شکل ملاحظه می‌شود، دو نمودار روش پیشنهادی دقیقا بر هم منطبق هستند و دلیل آن نبود کلاس جدید در داده‌های مجموعه داده Random RBF Generator Drift است. روش HAT در ابتدای این مجموعه داده به خوبی توانسته رده‌بندی داده‌ها را انجام دهد. ولی با گذشت زمان و زیاد شدن تغییر مفاهیم در یک بازه زمانی نسبتا زیاد نتوانسته تغییرات بوجود آمده را شناسایی کند و به همین دلیل افت شدیدی در داده‌های ۷۰۰۰ تا ۱۵۰۰۰ دارد. پس از آن با شناسایی بخشی از تغییرات دوباره مقدار صحت رده‌بندی افزایش می‌یابد، ولی با توجه به تغییر زیاد داده‌ها در مجموع نتوانسته میزان صحت قابل قبولی ارائه دهد.



شکل ۴-۵.نمودار تغییر صحت برای مجموعه داده Random RBF Generator Drift

در جدول ۴-۴ صحت روش پیشنهادی در مقایسه با سایر روش‌ها بر روی مجموعه داده‌های مختلف نشان داده شده است. مجموعه داده گوسی مدل ۱ همانطور که در ابتدای فصل مطرح گردید، یک مجموعه داده ساختگی بوده که فقط تغییر مفهوم در داده‌ها ایجاد شده است. در مجموعه داده گوسی مدل ۲ فقط کلاس جدید در نظر گرفته شده و برای مجموعه داده گوسی مدل ۳ علاوه بر تغییر مفهوم، کلاس جدید نیز در داده‌ها لحاظ شده است. همانطور که مشاهده می‌شود، در تمامی مجموعه داده‌ها روش پیشنهادی از صحت بالاتری نسبت به سایر روش‌ها برخوردار بوده و عملکرد بهتری دارد. این نشان می‌دهد که اکثر تغییرات موجود در داده‌ها از نوع تدریجی و ناگهانی است و این روش به خوبی توانسته تغییرات موجود در مجموعه داده‌های مذکور را تشخیص داده و با به روز رسانی مدل گروهی رده‌بند، در بهبود دقت و صحت الگوریتم موثر باشد. لازم به ذکر است که برای مجموعه داده Forest Cover، از داده ۱۸۰۰۰۰ تا ۲۸۰۰۰۰ برای اجرا انتخاب شده است. به همین دلیل مقدار صحت گزارش شده برای روش SENC-MaS با مقدار ثبت شده در مقاله مذکور متفاوت است. در مقاله میزان صحت ۷۹/۲ درصد ثبت شده که این مقدار برای کل مجموعه داده Forest Cover می‌باشد.

جدول ۴-۴.مقایسه صحت روش پیشنهادی و سایر روش‌ها بر روی مجموعه داده‌های مختلف

نام مجموعه داده	روش پیشنهادی بدون تشخیص کلاس جدید	روش پیشنهادی دارای تشخیص کلاس جدید	OBA	HAT	SENC-MaS
Forest Cover	۷۴/۶۷	**۷۵/۳۳**	۷۰/۱۶	۶۸/۹۸	۷۱/۰۵
Shuttle	۸۹/۶۱	**۹۰/۲۱**	۸۶/۵۷	۸۴/۲۶	۸۷/۷۹
Skin Segmentation	**۱۰۰**	**۱۰۰**	۹۹/۴	۹۹/۲۱	**۱۰۰**
Gaussian Generator model 1	**۹۸/۳**	**۹۸/۳**	۹۵/۲	۹۵/۱	۹۷/۰۴
Gaussian Generator model 2	۷۳/۹۶	**۹۵/۴۳**	۷۰/۷۴	۶۸/۱۹	۹۳/۲۸
Gaussian Generator model 3	۷۰/۰۳	**۹۵/۱۷**	۶۹/۵۱	۶۶/۹۴	۹۲/۷۳
Random RBF Generator Drift	**۸۱/۳۲**	**۸۱/۳۲**	۷۹/۸۳	۷۸/۰۶	۸۰/۰۷

۴-۳ جمع‌بندی

در این فصل به معرفی و بررسی آزمایش‌های متفاوت سیستم پیشنهادی پرداخته و سپس به مقایسه نتایج حاصل از آن‌ها بر روی مجموعه داده‌های مختلف اقدام کردیم. همانطور که مشاهده شد، راهکار پیشنهادی در تمامی آزمایش‌ها، عملکرد مناسبی را نسبت به رویکردهای قابل قیاس خود، به همراه داشت. این روش پیشنهادی که با استفاده از دو تکنیک متفاوت، تغییر مفاهیم و کلاس جدید را شناسایی می‌کند، به خوبی و با دقت بالایی کار رده‌بندی جریان داده را انجام می‌دهد. این روش با تشخیص به موقع تغییر مفهوم و کلاس جدید، تاثیر کم شدن دقت رده‌بندی به دلیل وجود این دو چالش را تا حد زیادی خنثی کرده و روش مناسبی برای رده‌بندی جریان داده محسوب می‌شود.

فصل ۵ : نتیجه گیری و پژوهش های آینده

۵-۱ نتیجه‌گیری

آنچه در این پایان‌نامه اشاره شد، معرفی و پیاده‌سازی یک رده‌بند گروهی است که بتواند تغییر و تکامل مفاهیم را در جریان داده تشخیص داده و با به روز رسانی مدل دقت رده‌بندی را بهبود بخشد. با توجه به اینکه در جریان داده برچسب همه داده‌ها در دسترس نیست، روش یادگیری نیمه‌نظارتی نسبت به روش نظارتی کارایی بیشتری دارد. گروهی بودن رده‌بند موجب می‌گردد بتوان علاوه بر داده‌های اولیه، از داده‌های جریان نیز برای رده‌بندی بهره برد. بنابراین با استفاده از داده‌های جدید مدل پیوسته به روز رسانی می‌شود. مدل ارائه شده از یک روش یادگیری نیمه‌نظارتی برای رده‌بندی جریان داده استفاده کرده که این روش علاوه بر تشخیص تغییر مفهوم، مفاهیم جدید را نیز شناسایی می‌کند. این مدل از تغییر آنتروپی برای تشخیص تغییر مفاهیم استفاده کرده و می‌تواند دو نوع تغییر مفهوم تدریجی و ناگهانی را در جریان داده تشخیص دهد؛ به طوری که با وارد شدن هر داده به پنجره تغییر آنتروپی داده‌ها بررسی شده و درصورتی که از حد معینی بیشتر بود، تغییر مفهوم تشخیص داده می‌شود. همچنین روش پیشنهادی کلاس جدید را با شناسایی داده‌های دور افتاده‌ای که همبستگی قوی بین خودشان دارند، تشخیص می‌دهد. با وارد شدن هر داده به بافر ضریب q-NSC مربوط به داده‌های بافر محاسبه شده و در صورتی که مقدار مثبتی بدست آید، کلاس جدید شناسایی می‌گردد.

در این پژوهش رده‌بند بر روی مجموعه داده‌های مختلفی با دو روش OBA و HAT مقایسه شده است. نتایج بدست آمده نشان دهنده عملکرد بهتر روش پیشنهادی نسبت روش‌های دیگر بر روی این مجموعه داده‌ها است.

۵-۲ پژوهش‌های آینده

روش پیشنهادی برای تشخیص تغییر مفاهیم از پنجره استفاده کرده و با محاسبه آنتروپی تنها یک تغییر مفهوم را به ازای داده‌های موجود در پنجره شناسایی می‌کند. زمانی که تغییر تشخیص داده می‌شود، داده‌های پنجره حذف شده و تا پر شدن پنجره تغییر مفهومی شناسایی نمی‌گردد. همچنین در صورتی که چند داده دارای تغییر مفهوم پشت سرهم بیایند فقط یکی از آن‌ها شناسایی می‌شود. بنابراین شناسایی چند داده دارای تغییر مفهوم به صورت پشت سرهم، از اقدامات ما برای انجام کارهای آینده است.

همانطور که عنوان شد، این روش تنها دو نوع تغییر مفهوم ناگهانی و تدریجی را تشخیص می‌دهد. بنابراین از موارد دیگری که می‌توان به عنوان پژوهش‌های آینده در نظر گرفت، اضافه شدن تشخیص تغییر مفهوم افزایشی و تکرار شونده است. تغییر مفهوم تکرار شونده به این دلیل در روش پیشنهادی قابل تشخیص نیست که طول پنجره به اندازه محدودی بوده و برای تکرار مفاهیم بسیار کوچک است.

به عنوان موردی دیگر می‌توان به تشخیص دو کلاس جدید پشت سرهم اشاره کرد. زمانی که دو کلاس جدید پشت سرهم یا با فاصله نزدیک به هم بیایند، فقط یکی از آن‌ها شناسایی می‌گردد. همچنین پس از تشخیص کلاس جدید و حذف داده‌های بافر، زمان زیادی برای پر شدن مجدد بافر سپری می‌شود که می‌تواند در تشخیص کلاس جدید؛ زمانی که تعداد کلاس‌های جدید پدیدار شده زیاد باشد، مشکل ایجاد کند.

ضمنا یکی دیگر از پژوهش‌های آینده پویا کردن اندازه پنجره و بافر است. به بیان دیگر، امکانی ایجاد شود که با اثبات روابط ریاضی بتوانیم اندازه پنجره و بافر را بدست آورده و با توجه به جریان داده‌های مختلف، متناسب با مقتضیات آن جریان، یک اندازه استاندارد مناسب برای پنجره و بافر مورد نظر تعیین کنیم.

مراجع

[1] Aggarwal, C. C. (Ed.). (**2007**). *Data streams:models and algorithms* (Vol. 31). Springer Science & Business Media.

[2] Masud, M. M., Chen, Q., Khan, L., Aggarwal, C., Gao, J., Han, J., & Thuraisingham, B. (**2010**, December). Addressing concept-evolution in concept-drifting data streams. In *2010 IEEE International Conference on Data Mining* (pp. 929-934). IEEE.

[3] Parker, B. S., & Khan, L. (**2015**, February). Detecting and tracking concept class drift and emergence in non-stationary fast data streams. In *Twenty-ninth AAAI conference on artificial intelligence*.

[4] Klinkenberg, R. (**2004**). Learning drifting concepts: Example selection vs. example weighting. *Intelligent data analysis*, *8*(3), 281-300.

[5] Bifet, A., & Gavalda, R. (**2007**, April). Learning from time-changing data with adaptive windowing. In *Proceedings of the 2007 SIAM international conference on data mining* (pp. 443-448). Society for Industrial and Applied Mathematics.

[6] Haque, A., Khan, L., & Baron, M. (**2016**, February). Sand: Semi-supervised adaptive novel class detection and classification over data stream. In *THIRTIETH AAAI Conference on Artificial Intelligence*.

[7] Kuncheva, L. I., & Faithfull, W. J. (**2013**). PCA feature extraction for change detection in multidimensional unlabeled data. *IEEE transactions on neural networks and learning systems*, *25*(1), 69-80.

[8] Sidhu, P., & Bhatia, M. P. S. (2018). A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority. *International Journal of Machine Learning and Cybernetics*, *9*(1), 37-61.

[9] Mahdi, O. A., Pardede, E., & Cao, J. (**2018**, January). Combination of information entropy and ensemble classification for detecting concept drift in data stream. In *Proceedings of the Australasian Computer Science Week Multiconference* (p. 13). ACM.

[10] Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (**1996**, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).

[11] Zhu, X., & Goldberg, A. B. (**2009**). Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, *3*(1), 1-130.

[12] Tsymbal, A. (**2004**). The problem of concept drift: definitions and related work. *Computer Science Department, Trinity College Dublin*, *106*(2), 58.

[13] Žliobaitė, I. (**2010**). Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*.

[14] Bifet, A., Holmes, G., Pfahringer, B., Kirkby, R., & Gavaldà, R. (**2009**, June). New ensemble methods for evolving data streams. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 139-148). ACM.

[15] Morshed, S. J., Rana, J., & Milrad, M. (**2016**, June). Real-time Data analytics: An algorithmic perspective. In *International Conference on Data Mining and Big Data* (pp. 311-320). Springer, Cham.

[16] Bifet, A., Holmes, G., Kirkby, R., & Pfahringer, B. (**2010**). Moa: Massive online analysis. *Journal of Machine Learning Research*, *11*(May), 1601-1604.

[17] Pfahringer, B., Holmes, G., & Kirkby, R. (**2008**, May). Handling numeric attributes in hoeffding trees. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 296-307). Springer, Berlin, Heidelberg.

[18] Mining, W. I. D. (**2006**). Data mining: Concepts and techniques. *Morgan Kaufinann*.

[19] Almalki, E. H., & Abdullah, M. (**2018**). Retracted: A survey on big data stream mining. *Journal of Fundamental and Applied Sciences*, *10*(4S), 278-284.

[20] Abdallah, Z. S., Gaber, M. M., Srinivasan, B., & Krishnaswamy, S. (**2016**). Anynovel: detection of novel concepts in evolving data streams. *Evolving Systems*, *7*(2), 73-93.

[21] Mansalis, S., Ntoutsi, E., Pelekis, N., & Theodoridis, Y. (**2018**). An evaluation of data stream clustering algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, *11*(4), 167-187.

[22] Nguyen, H. L., Woon, Y. K., & Ng, W. K. (**2015**). A survey on data stream clustering and classification. *Knowledge and information systems*, *45*(3), 535-569.

[23] Giannella, C., Han, J., Pei, J., Yan, X., & Yu, P. S. (**2003**). Mining frequent patterns in data streams at multiple time granularities. *Next generation data mining*, *212*, 191-212.

[24] Domingos, P., & Hulten, G. (**2000**, August). Mining high-speed data streams. In *Kdd* (Vol. 2, p. 4).

[25] Hulten, G., Spencer, L., & Domingos, P. (**2001**, August). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-106). ACM.

[26] Bach, S., & Maloof, M. (**2010**). A bayesian approach to concept drift. In *Advances in neural information processing systems* (pp. 127-135).

[27] Fearnhead, P., & Liu, Z. (**2007**). On-line inference for multiple changepoint problems. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *69*(4), 589-605.

[28] Ghazikhani, A., Monsefi, R., & Yazdi, H. S. (**2014**). Online neural network model for non-stationary and imbalanced data stream classification. *International Journal of Machine Learning and Cybernetics*, *5*(1), 51-62.

[29] Tsang, I. W., Kocsor, A., & Kwok, J. T. (**2007**, June). Simpler core vector machines with enclosing balls. In *Proceedings of the 24th international conference on Machine learning* (pp. 911-918). ACM.

[30] Rai, P., Daumé, H., & Venkatasubramanian, S. (**2009**, June). Streamed learning: one-pass SVMs. In *Twenty-First International Joint Conference on Artificial Intelligence*.

[31] Aggarwal, C. C., Han, J., Wang, J., & Yu, P. S. (**2006**). A framework for on-demand classification of evolving data streams. *IEEE Transactions on Knowledge and Data Engineering*, *18*(5), 577-589.

[32] Oza, N. C. (**2005**, October). Online bagging and boosting. In *2005 IEEE international conference on systems, man and cybernetics* (Vol. 3, pp. 2340-2345). Ieee.

[33] Abdualrhman, M. A. A., & Padma, M. C. (**2019**). CD2A: Concept Drift Detection Approach Toward Imbalanced Data Stream. In *Emerging Research in Electronics, Computer Science and Technology* (pp. 597-612). Springer, Singapore.

[34] Sheikhpour, R., Sarram, M. A., Gharaghani, S., & Chahooki, M. A. Z. (**2017**). A survey on semi-supervised feature selection methods. *Pattern Recognition*, *64*, 141-158.

[35] Chapelle, O., Scholkopf, B., & Zien, A. (**2009**). Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, *20*(3), 542-542.

[36] Bazan, J. G., Nguyen, H. S., Nguyen, S. H., Synak, P., & Wróblewski, J. (**2000**). Rough set algorithms in classification problem. In *Rough set methods and applications* (pp. 49-88). Physica, Heidelberg.

[37] Aggarwal, C. C. (**2015**). *Data mining: the textbook*. Springer.

[38] Nishida, K., & Yamauchi, K. (**2007**, October). Detecting concept drift using statistical testing. In *International conference on discovery science* (pp. 264-269). Springer, Berlin, Heidelberg.

[39] Nishida, K., Yamauchi, K., & Omori, T. (**2005**, June). ACE: Adaptive classifiers-ensemble system for concept-drifting environments. In *International Workshop on Multiple Classifier Systems* (pp. 176-185). Springer, Berlin, Heidelberg.

[40] Mu, X., Zhu, F., Du, J., Lim, E. P., & Zhou, Z. H. (**2017**, February). Streaming classification with emerging new class by class matrix sketching. In *Thirty-First AAAI Conference on Artificial Intelligence*.

[41] Gama, J. (**2010**). *Knowledge discovery from data streams*. Chapman and Hall/CRC.

[42] Meng, X., Bradley, J., Yavuz, B., Sparks, E., Venkataraman, S., Liu, D., ... & Xin, D. (**2016**). Mllib: Machine learning in apache spark. *The Journal of Machine Learning Research*, *17*(1), 1235-1241.

[43] Masud, M., Gao, J., Khan, L., Han, J., & Thuraisingham, B. M. (**2010**). Classification and novel class detection in concept-drifting data streams under time constraints. *IEEE Transactions on Knowledge and Data Engineering*, *23*(6), 859-874.

فهرست واژگان

ا

آنتروپی مشترک..... Joint entropy

افزایشی..... Incremental

ایستا..... Static

ب

برخط..... Online

برون‌خط..... Offline

پ

پیش‌فرض خمینه..... Manifold Assumption

پیش‌فرض خوشه..... Cluster Assumption

پیش‌فرض همواری ..... Smoothness assumption

پسین..... Posterior

پنجره‌گذاری..... Windownig

پنجره‌گذاری از یک نقطه مشخص..... LandMark window

پنجره‌های لغزان..... Sliding window

پنجره‌های محو شونده..... Fading window

پنجره‌های متمایل..... Tilted time window

پویا..... Dynamic

ت

تدریجی..... Gradual

تغییر مفهوم..... Concept drift

تکامل مفهوم..... Concept evolution

تکرارشونده..... Recurring

تمایزی..... Discriminative

تجزیه مقادیر منفرد..... Singular value decomposition (SVD)

تجزیه و تحلیل برخط حجم عظیم..... Massive Online Analysis(MOA)

تکنیک تشخیص تغییر صریح..... Explicit concept drift technique

ج

جریان داده..... Data Stream

جدایی..... Seperation

ح

حد آستانه..... Threshold

حسگرهای بی‌سیم..... Wireless sensor

خ

خوشه‌بندی..... Clustering

د

داده‌های خام..... Raw data

داده‌های دور افتاده..... Outlier

درخت تصمیم‌گیری خیلی سریع سازگار با مفهوم..... Concept-Adaptive Very Fast Decision Tree(CVFDT)

درخت تصمیم‌گیری خیلی سریع..... Very fast decision tree

درخت هافدینگ..... Hoeffding tree

دقت..... Percision

درخت تصمیم..... Decision tree

درست‌نمایی بیشینه..... Maximum likelihood

ر

رده‌بندی. ..... Classification

رده‌بند بیزین ..... Bayesian classification

رده‌بندی گروهی سازگار با تغییر مفهوم..... Adaptive classifiers ensemble system for concept drifting (ACE)

ردهبندهای گروهی..... Ensemble Classifiers

ش

شباهت..... Similarity

شبکه‌های عصبی..... Neural networks

ص

صحت..... Accuracy

ف

فراموشی تدریجی..... Gradual forgetting

فراموشی مفهوم..... Concept forgetting

فاصله اقلیدسی..... Euclidean distance

ق

قطعه داده..... Chunk

گ

گره..... Node

م

ماشین بردار پشتیبان..... Support Vector Machine(SVM)

ماشین بردار هسته..... Core Vector Machine

مفهوم جدید..... Novel concept

متن باز..... Open source

مدل خودآموز..... Self-training model

مدل آموزش مشترک..... Co-training model

مدل‌های مبتنی بر گراف..... Graph-based models

مدل‌های ترکیبی مولد..... Generative mixture models

مدل مقایسه‌ای بیزین ..... Bayesian Model Comparison (BMC)

ن

نزدیکترین همسایه..... K-Nearest neighbor(K-NN)

ناگهانی..... Sudden

ی

یادآوری..... Recall

یادگیری دسته‌ای..... Batch learning

Abstract

Data stream is a sequence of data generated from various information sources at a high speed and high volume. Among the main challenges of data stream analysis is the occurrence of concept drift. Concept drift is the change in statistical properties of data. Many available studies, to cope with the challenges of unlimited data stream length and also concept drift, use approaches assuming existence of true labels for all data. Nevertheless, regarding the cost of labeling instances, it is often assumed that only a part of instances are labeled. Another important challenge in data stream analysis is concept evolution. When the change of data leads to the emergence of new concepts, the evolution of the concept occurs and appears as a new emerging class. In this paper, a semi-supervised ensemble learning algorithm is proposed which uses entropy variation to detect concept drift in data stream classification. The proposed ensemble learning model is trained with a limited initial labeled set. Afterwards, in occurrence of concept drift, the unlabeled data is used to update the ensemble model. Also, in this method, the new class is identified by detected outliers with a strong correlation among them. The proposed method is capable of detecting changes in data, and improve its accuracy via updating the learning model. Experiment results show that the proposed method is more effective than other methods in different aspects.

Keywords: Data stream, ensemble learning, Concept drift, Entropy, Semi-supervised classification, Concept Evolution

Shahrood University of Technology

Faculty of Computer Engineering

M.Sc Thesis in Artificial Intelligence Engineering

Semi-supervised classification of evolving data streams

By: : Hossein Hasan Nezhad Namaghi

Supervisor:

Dr. Hoda Mashayekhi

Advisor:

Dr. Morteza Zahedi

July 2019