ششمین همایش بین‌المللی افق های نویـن در
مهندسی برق، کامپیوتـر و مکانیک
6th International Conference on the New Horizons in
Electrical Engineering, Computer and Mechanical
www.mhconf.ir

دسته‌بندی روش های شناسایی رانش مفهوم

نگارنده: امیر کتول

دانشجوی کارشناسی ارشد دانشگاه گلستان گرگان (گرایش نرم افزار)amirkatool@gmail.com

چکیده

بخش عظیمی از دنیای تکنولوژی امروز را سیستم هایی تشکیل داده اند که با جریانی پیوسته و سـریع از داده کـار مـی‌کننـد(جریان داده). بحث یادگیری در جریان‌داده با چالش‌های مختلفی از جمله دسترسی محدود به داده ها و تغییر توزیع داده ها در گذر زمـان(رانش مفهوم) روبه‌رو است.در طول دهه اخیر توجه به بحث رانش‌مفهوم بیشتر شده و شناسایی آن به مسئله محوری در یادگیری آنلایـن و داده کاوی از جریان‌داده تبدیل شده است .بسیاری از روش های شناسایی رانش‌مفهوم براساس کنترل عملکرد کلاس‌بندی‌کننده و پیش‌بینـی کلاس حقیقی داده ها کار می‌کنند.دسترسی به کلاس حقیقی داده ها و میزان تاخیر این دسترسی عاملی تعیین کننده در میزان کارایی روش های شناسایی رانش‌مفهوم در شرایط مختلف است. بر همین اساس، دسته‌بندی برای روش های شناسـایی رانـش مفهـوم در سـال ۲۰۱۵ صورت گرفته است که به سه دسته روش های نظارتی،بدون نظارت و نیمه نظارتی تقسیم می‌شود.با توجه به اینکه این دسته‌بنـدی شامل روش‌های مطرح شده بعد از سال ۲۰۱۵ نمی باشد؛در این گزارش سعی شده تا با پر کردن این جای خالی و تکمیل این دسته‌بندی با روش های شناسایی رانش‌مفهوم مطرح شده بعد از سال۲۰۱۵ و معرفی اجمالی هرکدام از آنها، قدمی مثبت در تفکیک این روش هـا از هم برداشته شود.

کلمات کلیدی:جریان‌داده، یادگیری ماشین، رانش مفهوم، روش‌های شناسایی نظارتی،بدون نظارت،نیمه نظارتی

۱- مقدمه

در دنیای امروز سرعت رشد و توسعه تکنولوژی فراتر از تصورات و پیش‌بینی های گذشته است.تخمین حجم اطلاعات تولیـدی روزانـه در دنیای امروز کاری تقریبا نشدنی است.داده،در هر لحظه،در بسترهای مختلف با سرعت بالا تولید و منتشر شده و توانایی تحلیل و یادگیری از داده در این شرایط را به چالشی جدید برای دنیای تکنولوژی تبدیل می‌کند.یادگیری از این جریان بی اندازه داده که به صورت پیوسـته و پی‌درپی تولید و ارسال می‌شود و آن را با نام جریان‌داده می‌شناسیم(Data Stream)،چالش‌های خاص خود را دارد. از چالش‌های اصلی برای یادگیری در جریان‌داده، حجم و اندازه غیرقابل اندازه‌گیری و بیش از اندازه داده‌ها؛ سرعت انتشار بسیار بالا و دسترسـی محـدود بـه داده ها که در بسیاری از شرایط تنها یکبار قابل دسترسی هستند و تغییر داده ها در گذر زمان(رانش مفهوم)؛ را می توان نام بـرد. اصـلی ترین چالش در یادگیری جریان‌داده،شناسایی رخداد رانش مفهوم است.رانش مفهوم که به تغییر ویژگی‌های آماری متغیر مورد مطالعه بـا گذر زمان در این محیط‌های پویا گفته می‌شود،اصلی ترین دلیل افت کیفی سیستم های مبتنی بر استخراج دانش در گذر زمـان شـناخته شده است.هرچند توجه به رانش مفهوم وتلاش برای شناسایی آن در سال های اخیر بیشتر شده است،اما هنوز ایـن مبحـث در قـدم‌هـای اولیه خود قرار داشته و ناشناخته‌های زیادی پیرامون آن وجود دارد.بیشتر روش‌های شناسایی رانش‌مفهوم با توجـه بـه بررسـی عملکـرد کلاس‌بندی‌کننده کار می‌کنند و ضعف اصلی آن‌ها این است که فرض می‌کنند که کلاس حقیقی داده ها در هر لحظه قابل دسترسی بوده که در بسیاری از شرایط دنیای حقیقی امری ناممکن است.در سال‌های اخیر روش‌های جدیدی با در نظرگیری تغییرات ابعاد ویژگی‌هـای داده،برای شناسایی رانش‌مفهوم مطرح شده است.مقالات متعددی نیز بر اساس معیار های مختلف،دسته‌بندی‌هایی روی این روش‌ها انجام داده اند اما مقاله ای که تقسیم بندی خود را براساس نظارتی یا بدون نظارت بودن،با توجه به جدید ترین روش‌ها(۲۰۱۵ بـه بعـد) انجـام دهد،تا آنجا که در دانش این گزارش می‌گنجد صورت نپذیرفته است.در این گزارش بعد از تعریف و بررسی رانش‌مفهوم،سعی شده تا جدید ترین روش‌های شناسایی رانش مفهوم در یک دسته‌بندی بر اساس نظارتی یا نیمه نظـارتی و بـدون نظارت،دسـته‌بنـدی شود.ضـمنا ایـن روش‌ها به صورت اجمالی بررسی و معرفی شوند.

۲-رانش مفهوم

۲,۱-تعریف

به پدیده تغییر توزیع داده در گذر زمان رانش مفهوم(Concept Drift)می‌گویند.در جریان‌داده،بسیاری مواقع،داده‌های جدید، از الگـو ها جدیدی پیروی می‌کنند که در داده‌های پیشین مشاهده نشده است.این تغییرات به دلایل پیش‌بینی نشده‌ای ممکن اسـت رخ دهد.بـه عنوان مثال در یک سیستم تشخیص کلاهبرداری،یک گونه جدید از کلاهبرداری‌های کارت اعتباری ممکن است مشاهده شود که پیشـتر وجود نداشته است.با این تغییرات داده،الگوها و کلاس‌بندی‌کننده‌هایی که قبل‌تر و با داده‌های قبل از این تغییرات ایجاد شـده انـد، دچـار افت عملکرد خواهند شد.در تعریف رانش مفهوم داریم:

در بازه زمانی صفر تا t،یک مجموعه از نمونه‌های $S_{0,t} = \{d_0, \dots, d_t\}$ داریم که $d_i = (X_i, y_i)$ است.X_i بردار ویژگـی‌هـای داده و y_i کلاس مربوط به آن است.رانش مفهوم بدین صورت تعریف می‌شود که در طور گذر زمان توزیع داده دچار تغییراتی می‌شود که در اثر ایـن تغییرات خواهیم داشت $\exists t: p_t(X, y) \neq p_{t+1}(X, y)$ احتمال مشترک بین X و y است.رانش مفهوم پدیده‌ای پیچیده است که به شکل های مختلفی می‌تواند رخ دهد که بررسی آن می‌پردازیم.

۲,۲-انواع رانش مفهوم

اگر رابطه احتمال مشترک را براساس رابطه بیز بنویسیم خواهیم داشت: $p_t(X, y) = p_t(X) \times p_t(y|X)$.

براساس این رابطه تغییرات احتمال مشترک می‌تواند ناشی از تغییرات:

الف) $p_t(X)$ باشد.در اثر این تغییرات توزیع موجود درون کلاس‌های تعیین شده داده دچار تغییر شده اما این تغییرات اثری روی مرزهای کلاس‌بندی نداشته و عملکرد کلاس‌بندی‌کننده را تضعیف نمی‌کند و به آن، رانش مجازی نیز گفته می‌شود.

ب)$p_t(y|X)$ باشد.در اثر این تغییرات،کلاس تعیین شده برای مجموعه ویژگی‌های معین دچار تغییر شده و به دنبال آن مرزهای کلاس-بندی هم دچار تغییر می‌شود.به آن رانش واقعی گفته شده و در اثر آن عملکرد کلاس‌بندی‌کننده افت خواهد کرد.

لازم به ذکر است که در بسیاری از پدیده های واقعی هر دو حالت الف و ب مـی‌توانـد بـه صـورت همزمـان رخ دهـد.همچنین دو پدیـده پدیداری کلاس جدید(Novel Class Emergence) و ادغام کلاس‌هـای موجـود(Existing Class Fusion) کـه در گذشـته بـه عنـوان گونه‌ای مجزا از رانش‌مفهوم در نظر گرفته نمی‌شد را می توان به عنوان دسته سوم انواع رانش که در اثر تغییرات احتمال پیشـین $p_t(y)$ رخ میدهد،در نظر گرفت.

رانش مفهوم پدیده پیچیده ای هست که از لحاظ سرعت،شدت وقابل پیش‌بینی بودن نیز به انواع مختلفـی تقسـیم شـده و مـی‌توانـد بـه شکل‌های مختلفی رخ دهد.این تنوع تشخیص رانش‌مفهوم را سخت تر کرده و باعث می‌شود بسیاری از روش ها تنهـا در شناسـایی گونـه هایی خاص ممتاز باشند. Khamassi, Sayed-Mouchaweh et al. (2018)

۳-شناسایی رانش مفهوم

۳,۱-مطالعات پیشین

در سال های اخیر مقالات متعددی در رابطه با بحث روش‌های شناسایی رانش‌مفهوم مطرح شده است.در این قسمت سـعی شـده تـا آخرین مقالات جامع و مروری پیرامون روش‌های شناسایی رانش‌مفهوم بررسی و به صورت مختصر بیان شوند.

در مقاله Lu, Liu et al. (2018) چارچوبی کلی برای شناسایی رانش‌مفهوم ارائه شده اسـت کـه متشـکل از چهـار مرحلـه مختلـف است:بازیابی داده(Data Retrieval)،مدل سازی داده(Data Modeling)،محاسبه آزمون‌هـای آمـاری(Test Statistics Calculation) و آزمون فرضیه(Hypothesis Test).با توجه به چارچوب مطرح شده تقسیم روش‌های مختلف شناسایی رانش مفهوم در این مقاله براساس تکنیک های مختلفی است که در مرحله سوم یعنی محاسبه آزمون آماری استفاده می‌شود.این مقاله روش های شناسایی رانش مفهـوم را به سه دسته کلی تقسیم می‌کند.دسته اول روش هایی است که براساس کنترل نرخ خطا کلاس‌بندی‌کننده کار می‌کنند.دسته دوم شـامل روش‌هایی است که از اندازه‌گیری‌های فاصله بین داده‌های جدید و قدیم برای تخمین میزان تغییرات استفاده می‌کنند و مبتنی بر کنتـرل توزیع داده هستند.دسته سوم شامل روش‌هایی است که از چندین آزمون فرضیه با هم برای شناسایی رانش مفهوم استفاده می‌کننـد کـه خود به دو دسته موازی و سری تقسیم می‌شود.

در مقالـه Hu, Kantardzic et al. (2020) دسـته‌بنـدی روش‌هـای شناسـایی رانـش مفهـوم بـه دو دسـته کلـی مبتنـی بـر عملکرد(performance-based)و مبتنی بر توزیع داده(data distribution-based) تقسیم شده است.روش‌های مبتنـی بـر عملکـرد بـه صورت پیوسته،معیاری مربوط به میزان خطا مانند درستی(accuracy)،میزان اف(F-measure)، دقت (precision) و بازیـابی(recall) را کنترل می‌کنند.به صورت کلی در صورتی که افت محسوسی در این معیارها رخ دهد خبر از رخداد رانش مفهوم می‌دهد.روش‌های مبتنـی بر توزیع داده بر اساس کنترل تغییر معیارهای توزیع داده مانند مکان(location)،چگالی (density) و دامنه(range) کار می‌کنند.لازم بـه ذکر است روش‌های موجود در این دسته با روش‌های دو دسته آخر مقاله Lu, Liu et al. (2018) همپوشانی دارند.

در مقاله Wares, Isaacs et al. (2019) دسته‌بندی مختص روش های نظارتی مطرح می‌شود.این دسته‌بندی به چهار گروه:روش‌های آماری(Statistical methods)،روش‌های مبتنی بر پنجره(window-based methods)،روش‌های دسته جمعی مبتنی بر بلـوک(block-based ensembles)و روش‌های دسته‌جمعی افزایشی(incremental-based ensembles) تقسیم می‌شوند.دسته اول،شـامل روش‌هـایی است که از آزمون های آماری نظیر Page–Hinckley برای شناسایی رانش استفاده میکنند.دسته دوم شامل روش‌هایی است که براساس درستی کلاس‌بندی‌کننده برای نمونه های درون یک پنجره کار می‌کنند.دو دسته آخر براساس درستی کلاس‌بندی‌کننده دسته‌جمعی کار می‌کنند و تفاوت آن‌ها بر اساس نحوه واکنش به رانش مفهوم است.دسته سوم نمونه‌ها را به صورت بلوک داده برای آموزش مجدد استفاده می‌کنند در حالی که دسته آخر با رسیدن هر داده جدید به صورت افزایشی مدل را بازآموزی می‌کند.

در مقاله(2020) Gemaque, Costa et al. دسته‌بندی براساس نحوه عملکرد روش شناسایی رانش مفهوم مخـتص روش‌هـای نیمـه نظارتی و بدون نظارت انجام شده است.این دسته‌بندی روش‌های نیمه نظارتی و بدون نظارت را بر اساس اینکه داده هـا را بـا چـه قـالبی بررسی می‌کنند به دو دسته روش‌های شناسایی رانش مفهوم بسته‌ای(batch drift detection methods)و روش‌هـای شناسـایی رانـش مفهوم آنلاین(online drift detection methods) تقسیم می‌کند.روش‌های بسته ای داده ها را در قالب یک بسته از جریان داده جمـع-آوری کرده و سپس رخداد رانش آن را بررسی می‌کنند که خود بر اساس اینکه از تمام داده های یک بسته استفاده مـی‌کننـد یـا تنهـا از بخشی از آن به دو دسته تشخیص با کل بسته و تشخیص با جزئی از بسته تقسیم می شوند.در مقابل روش‌های شناسایی آنلاین با رسیدن هر داده رخداد رانش را بررسی می‌کنند و بر اساس اینکه از پنجره ای با اندازه ثابت یا متغیر برای مقایسه داده‌های جدید و قدیم اسـتفاده می‌کنند به دو دسته روش‌های آنلاین با پنجره ثابت و متغیر تقسیم می‌شوند.

در مقاله Khamassi, Sayed-Mouchaweh et al. (2018) دسته‌بندی کلی بر اساس پاسخ به پنج سوال صورت گرفتـه اسـت:داده چگونه پردازش می‌شود؟،یادگیری چگونه انجام می‌شود؟،رانش مفهوم چگونه نظارت می‌شود؟،رانش مفهوم چگونـه مـدیریت مـی‌شـود؟ و معیارهای عملکرد چه هستند؟.در پاسخ به سوال سوم مطرح شده در این مقاله که در این گزارش مد نظر ماست،انواع روش‌های شناسایی رانش مفهوم به سه دسته نظارتی،نیمه نظارتی و بدون نظارت براساس نیاز به کلاس حقیقی داده، تقسیم شده اند.

با توجه به اینکه دسته‌بندی که در مقاله Khamassi, Sayed-Mouchaweh et al. (2018) صورت گرفته است،شـامل بسـیاری از روش های نیمه نظارتی و بدون نظارت که پس از آن ارائه شده اند نیست،در این گزارش سعی شده تا براسـاس همـان روش دسـته‌بنـدی روش هایی که بعد از سال ۲۰۱۵ مطرح شده اند بررسی و دسته‌بندی شوند.با توجه به این دسته‌بندی، روش‌های دسته اول مقاله‌های Lu, Liu et al. (2018)وHu, Kantardzic et al. (2020) در دسته نظارتی قرار می‌گیرند.باقی روش های مطرح شده در این مقـالات،برخی در دسته نظارتی و برخی در دسته بدون نظارت قرار می‌گیرند که در ادامه به بررسی آن‌ها میپردازیم.

۳,۲-روش های نظارتی

روش‌های نظارتی روش‌هایی هستند که نتیجه پیش‌بینی کلاس هر داده را در هر زمان آماده فرض مـی‌کننـد.این روش‌هـا براسـاس کنترل معیارهای میزان خطا کلاس‌بندی کننده ماننـد دقت،درسـتی،بازیابی و...کـار مـی‌کننـد و کنتـرل تغییـرات P(Y|X) را بـر عهـده دارند.ساختار کار این دسته از روش‌های شناسایی رانش‌مفهوم،شرایط استقلال از روش یادگیری و درستی بالا در شناسایی همراه با امکان شناسایی انواع رانش مفهوم را ممکن می‌سازد.در مقابل، فرض وجود کلاس حقیقی هر داده در هر لحظه در بسیاری از شـرایط نـاممکن و در صورت وجود بسیار هزینه بر است.

روش FW-DDM(Fuzzy Windowing Drift Detection Method) در Liu, Zhang et al. (2017) یکی از روش‌هـای نظـارتی است که برخلاف روش‌های شناسایی رانش مفهوم نظارتی مرسوم که از پنجره استفاده می‌کنند،با استفاده از پنجره زمانی فازی و اجازه به اینکه دوره همپوشان از مفهوم جدید و قدیم در کنار هم در پنجره لغزنده قرار گیرند،توانست نسبت به روش‌های مرسـوم نظـارتی دقـت و درستی بهتری کسب کند.

روش DELM(Dynamic Extreme Learning Machine) کـه در Xu and Wang (2017) مطـرح شـد،این روش از ELM کـه یک روش شبکه عصبی با یک لایه پنهان است استفاده می‌کند.این روش از DDM که یکی از قدیمی تـرین روش هـای شناسـایی رانـش مفهوم نظارتی است(Gama, Medas et al. (2004 استفاده می‌کند تا رانش را شناسایی کند.بعد از اخطار رخداد رانش تعـداد گـره‌هـای لایه مخفی افزایش پیدا می‌کند تا کلاس‌بندی‌کننده عمومی‌تری یادگیری شود.

روش LFR (Linear Four Rate) که در Wang and Abraham (2015) مطرح شد،از روش‌های چند آزمون فرضیه مـوازی اسـت که با کنترل همزمان چهار معیارTP،FP،TNوFN رخداد رانش را کنترل می‌کند.با توجه به اینکه این روش به طور همزمان چهـار معیـار مختلف طبقه‌بندکننده را کنترل می‌کند حساسیت به تغییر آن بسیار زیاد است و به همین خاطر رخـداد رانـش مفهـوم را بـا نـرخ بـالا و همین طور بسیار سریع شناسایی می‌کند و در عین حال از مشکلات آن نرخ بالای False Alarm به خاطر همین حساسیت زیاد است.

یکی دیگر از روش‌های نظارتی چند آزمون فرضیه روش سه لایه(three-layer) مطرح شده در Zhang, Chu et al. (2017) است.این روش که برای شناسایی رانش مفهوم در متن در ابتدا مطرح شد،از سه لایه تشکیل شده است که براساس ارزش اطلاعات(Information Value) و شباهت ژاکار(Jaccard similarity) رانش مفهوم را شناسایی می‌کند.در لایه یک تغییرات کلاس حقیقی داده‌ها،P(y)؛در لایه دوم تغییرات فضای ویژگی‌ها،P(X)و در لایه سوم تغییرات رابطه بین ویژگی‌ها و کلاس حقیقی داده،P(y|X) بررسی می شود.در این روش با محاسبه ارزش اطلاعات و وزن شواهد(Weight of Evidence)بین دو بسته داده جدید و قدیمی میزان تغییرات محاسبه و مقایسه می‌شود.

روش Hierarchical Change-Detection Tests (HCDTs) که در Alippi, Boracchi et al. (2016) مطرح شد ساختار کلی روش‌های چند آزمون فرضیه سری را بیان می‌کند.این روش از دو لایه شناسایی(detection layer)و لایه تصدیق(validation layer)تشکیل شده است.در لایه شناسایی هر کدام از روش‌های موجود شناسایی رانش که تاخیر شناسایی پایین و نیاز محاسباتی کمتری دارند قابل استفاده است.بعد از شناسایی رانش در این لایه،لایه تصدیق که دارای محاسبات سنگین‌تر و پیچیده‌تری است و می توان از آزمون های فرضیه معروف در آن استفاده کرد،فعال شده و رخداد رانش یا عدم آن را تایید می‌کند.با وجود اینکه می توان از روش‌های مختلفی در دو لایه شناسایی و تصدیق استفاده کرد و این روش به شکل‌های مختلف مطرح می‌شود.چون نگارنده مقاله از تکنیک های آماری برای لایه تشخیص استفاده کرده است که نظارتی هستند؛ این روش در دسته نظارتی در این گزارش مطرح شده است.

روش Hierarchical Linear Four Rate (HLFR) مطرح شده در Yu and Abraham (2017) یکی دیگر از روش‌های چند آزمون فرضیه سری است.در این روش برای لایه شناسایی، روش LFR مورد استفاده قرار گرفته است.نرخ بالای امکان تشخیص اشتباه روش LFR،حال با وجود لایه تصدیق و وجود روشی برای تایید یا رد رخداد رانش، بهتر شده و امکان تشخیص انواع شکل‌های رانش را برای این روش ممکن می‌سازد.در لایه تصدیق از یک آزمون جایگشتی(Permutation Test) برای تایید رخداد رانش استفاده شده است.

در مقاله de Lima Cabral and de Barros (2018) روش، Nishida Statistical Test of Equal Proportions (STEPD) and Yamauchi (2007)که یکی از روش‌های معروف برای شناسایی رانش مفهوم نظارتی است؛ با استفاده از آزمون فیشر بهبود یافته است.روش مطرح شده از نظر اینکه دو پنجره از داده را با هم مقایسه کرده و در صورت تجاوز از حد آستانه اخطار رانش مفهوم را دهد مانند STEPD عمل می‌کند،با این تفاوت که با بکارگیری آزمون دقیق فیشر(Fisher’s Exact test) به جای اینکه براساس نرخ پیش-بینی درست کار کند،براساس تغییرات میزان خطا کار می‌کند.استفاده از این روش آزمون آماری، این امکان را بوجود می‌آورد که برخلاف روش STEPD که پیشتر برای شرایطی نظیر تعداد نمونه داده محدود یا داده نامتعادل (imbalanced)، که جوابگو نبود؛ اکنون جوابگو باشد.در بررسی های انجام شده،این روش توانست گونه‌های بیشتری از رانش را بهتر از روش STEPD شناسایی کند.

روش McDiarmid Drift Detection Method (MDDM) مطرح شده در Pesaranghader, Viktor et al. (2018) با استفاده از از نابرابری مکدیارمید(McDiarmid’s inequality) و وزن‌دهی به پیش‌بینی داده در پنجره لغزان، توانست تاخیر شناسایی رانش مفهوم را نسبت برخی روش‌ها نظیر Page-Hinckley بهبود ببخشد.در این روش به داده‌های جدید وزن بیشتری داده می‌شود تا اهمیت آنهادر محاسبات بیشتر شود.در پنجره لغزان میانگین وزن داده ها محاسبه شده و با ماکسیمم میانگین وزن موجود مقایسه می‌شود،در صورتی که میزان تفاوت از حدود نابرابری مکدیارمید بیشتر باشد،رخداد رانش تایید می‌شود.

روش Hoeffding Drift Detection Method (FHDDM) مطرح شده در Pesaranghader and Viktor (2016) نیز برای کاهش تاخیر شناسایی رانش مفهوم روشی مبتنی بر نابرابری هافدینگ(Hoeffding’s inequality) در پنجره لغزان ارائه می‌کند.این روش براساس مقایسه احتمال بیشینه پیش‌بینی درست کلی با احتمال پیش‌بینی درست داده‌های جدید،رخداد رانش را شناسایی می-کند.در صورتی که رانش رخ نداده باشد تفاوت محسوس نخواهد بود.

روش Multiscale Drift Detection Test (MDDT) مطرح شده در Wang, Kang et al. (2018) با نگاه کاهش حجم محاسباتی برای شناسایی رانش مفهوم، روشی مبتنی بر نمونه گیری مجدد(resampling) و آزمون تی(paired student's t test)ارائه

می‌کند.در این روش داده‌های قدیمی با نمونه گیری و به صورت تصادفی با داده‌های جدید برای تایید رخداد رانش مقایسه می‌شوند.کاهش حجم محاسبات برای شناسایی رانش باعث شده این روش برای شناسایی رانش بسیار سریع عمل کند.

بسیاری از روش‌های شناسایی رانش‌مفهوم در مدل های دست‌جمعی کلاس‌بندی‌کننده ها کار می‌کنند.روش Number and Distance of Errors(NDE) در Dehghan, Beigy et al. (2016) مطرح شد،روشی است که هر یک از نمونه‌ها را بررسی کرده و توزیع خطا مدل دست‌جمعی کلاس‌بندی را کنترل می‌کند.NDE تخمینی از میزان خطا کلاس‌بندی هر نمونه ارائه میدهد،در صورتی که NDE داده جدید با NDE کلی مجموعه تفاوت محسوسی داشته باشد،رخداد رانش اعلام می‌شود.

روش diversified dynamic-weighted majority(DDWM) که در Sidhu and Bhatia (2018) مطرح شده از دو مدل دسته‌جمعی کلاس‌بندی‌کننده استفاده می‌کند،یکی با تنوع پایین و دیگری با تنوع بیشتر،با اینکار روش پیشنهاد شده دقت بهتری برای کلاس‌بندی و شناسایی انواع مختلف رانش مفهوم پیدا می‌کند.به طور کلی اگر هر کدام در شناسایی رانش گونه‌ای از رانش مفهوم ضعیف عمل کنند،دیگری آنرا جبران خواهد کرد.هر مدل موجود در هر دو کلاس‌بندی‌کننده دسته‌جمعی(تنوع پایین و بالا)برای پیش‌بینی کلاس هر داده جدید تاثیر گذار خواهند بود.این روش با وزن‌دهی برای پیش‌بینی کلی هر دو مدل دسته‌جمعی و کلاس‌بندی‌کننده‌های پایه درون آنها کار می‌کند.با پیش‌بینی نادرست وزن‌ها کاهش می‌یابد و اگر از میزان آستانه ای عبور کند،آن کلاس‌بندی‌کننده پایه حذف می‌شود.در صورتی که پیش‌بینی کلی مجموعه اشتباه باشد،هر دو مدل دسته‌جمعی متنوع و کم تنوع کنترل شده و کلاس‌بندی کننده ضعیف حذف شده و کلاس‌بندی‌کننده جدیدی به مجموعه اضافه می‌شود.

در مدل کلاس‌بندی‌کننده های دسته‌جمعی،تکنیک های مختلفی برای ایجاد تنوع کلاس‌بندی‌کننده‌های پایه وجود دارد که هرکدام نقاط ضعف و قوتی دارند.در روش EnsembleEDIST2 مطرح شده در Khamassi, Sayed-Mouchaweh et al. (2019) با ترکیب سه روش مختلف تنوع مبتنی بر بلوک داده(block-based data)،وزن داده(weighting-data) و فیلتر کردن داده(filtering-data) سعی شده تا عملکرد هریک را در کنار دیگری بهبود داده و ضعف های آنرا جبران کند.این روش از EDIST2 که یک روش شناسایی رانش مفهوم است که با نظارت بر عملکرد مدل دسته‌جمعی(ensemble) رانش را شناسایی می‌کند؛ استفاده می‌کند.Khamassi, Sayed-Mouchaweh et al. (2015) .با ترکیب این سه روش ایجاد تنوع،این روش قادر است تا انواع مختلف رانش مفهوم را شناسایی کند.

روش EBE (Entropy Based Ensemble)که در Mahdi, Pardede et al. (2018) مطرح شده است از آنتروپی(entropy) برای شناسایی رانش مفهوم در کلاس‌بندی‌کننده دسته‌جمعی استفاده می‌کند.در این روش آنتروپی اطلاعات(information entropy) برای هریک از کلاس‌بندی‌کننده‌های پایه از یک بلوک داده با اندازه ثابت محاسبه می‌شود.کلاس‌بندی‌کننده ها براساس اینکه سطح آنتروپی آن‌ها مورد تایید است یا خیر در مدل دسته‌جمعی وارد یا از آن حذف می‌شوند.

بیشتر روش‌های دسته‌جمعی از روش رای‌گیری سنتی برای کلاس‌بندی کردن استفاده می‌کنند.در Krawczyk and Cano (2018) تغییری در روش رای‌گیری صورت گرفته است.در این روش امکان اینکه کلاس‌بندی کننده‌ها بتوانند از شرکت در رای گیری نهایی پرهیز کنند فراهم شده است. اطمینان(confidence) تمام کلاس‌بندی‌کننده‌ها برای هریک از نمونه ها کنترل شده و تنها کلاس‌بندی‌هایی که میزان اطمینان آن از میزان آستانه بیشتر باشد قادر به شرکت در رای‌گیری نهایی هستند.این میزان آستانه به صورت پویا از جریان داده بدست می‌آید.امکان خودداری از شرکت در رای‌گیری نهایی در این روش مقاوت این روش دست‌جمعی نسبت به نویز را افزایش داده است.نویسنده مقاله بیان می‌کند که تکنیک اعمال شده،تکنیکی کم هزینه و ساده است که می توان برای بهبود همه روش‌های دسته‌جمعی آنلاین مورد استفاده قرار گیرد.

برخی از روش‌ها از ابزار متفاوتی برای شناسایی رانش مفهوم استفاده می‌کنند.در Jaworski, Duda et al. (2017) از Restricted Boltzmann Machine (RBM) برای شناسایی رانش مفهوم استفاده می شود که خود گونه‌ای از روش‌های شبکه عصبی است.RBM قادر است تا مدلی بنیادی از داده را یادگیری کند. بعد از مرحله آموزش RBM حاوی اطلاعاتی فشرده شده از توزیع داده ای است که از آن یادگیری انجام داده است.نویسنده مقاله بیان می‌کند که می‌توان از این اطلاعات فشرده شده برای مقایسه یکسانی توزیع داده‌های

جدید و قدیم استفاده شود.دو شاخص انرژی آزاد(free energy) و خطـای بازسـازی(reconstruction error) بـرای محاسـبه و مقایسـه پیشنهاد شده است.روش دیگری که از شبکه عصبی برای شناسایی رانش مفهـوم اسـتفاده مـی کنـد در Lobo, Del Ser et al. (2018) مطرح شده است که از Evolving Spiking Neural Networks استفاده می‌کند که براساس تغییرات میزان ادغام نورون های خود مـی تواند رخداد رانش را شناسایی کند.این دو روش از جهت اینکه از ابتدا به شبکه عصبی مورد استفاده خود برای شناسـایی رانـش نیازمنـد بوده و در مرحله یادگیری خود در واقع شناسایی رانش‌مفهوم را شروع کرده اند که در آن ها کلاس حقیقی داده مورد نیاز است.به همـین خاطر در این گزارش در دسته نظارتی قرار گرفته اند.

روش Online Histogram-based Naïve Bayes Classifier (OHNBC) که در Astudillo, González et al. (2016) مطـرح شده است با استفاده از هیستوگرام‌های تک بعدی(uni-dimensional histograms) پیش‌بینی کلاس نمونه ها را انجام میدهدو با کنترل تغییرات عملکرد کلاس‌بندی رخداد رانش مفهوم را شناسایی می‌کند.در این روش آموزش و آزمون و استقرار کلاس‌بنـدی‌هـا بـه صـورت متناوب انجام شده و امکان اینکه این روش در جریان‌داده ای که در آن داده هایی با کلاس معین و نامعین درکنار هم وجود دارد جوابگـو باشد؛ را ممکن می‌کند.

۳,۳-روش‌های نیمه نظارتی

اهمیت توانایی فعالیت به صورت بدون نظارت باعث شده است تا بسیاری از روش‌ها با ترکیب روش‌های نظـارتی بـا روش‌هـای بـدون نظارت رانش مفهوم را شناسایی کنند.این دسته از روش‌ها در دسته نیمه نظارتی در این مطالعه قرار گرفته انـد.روش‌هـای نیمـه نظـارتی روش‌هایی هستند که به تعداد کمتری نمونه با کلاس حقیقی نیازمند هستند.این روش‌ها در شرایطی که کلاس حقیقی همـه داده هـا در دسترس نیست جوابگو خواهند بود.

روش knowledge-maximized ensemble (KME) کـه در Ren, Liao et al. (2018) مطـرح شـده اسـت از همـین تکنیـک استفاده می‌کند.در این روش از داده‌هایی با کلاس حقیقی معین و داده‌هایی بدون کلاس به صورت همزمان اسـتفاده مـی‌شـود تـا دانـش مربوط به مفهوم کنونی بیشتر شود.در این روش هم عملکرد کلاس‌بندی‌کننده و هم تغییرات توزیع نمونه ها کنترل می‌شـود.از اطلاعـات بدست آمده از داده‌های با کلاس حقیقی و بدون کلاس حقیقی،وزن اعضای مدل دسته‌جمعی محاسبه و کنترل می‌شود.با اینکه این روش از داده های بدون کلاس حقیقی نیز استفاده می‌کند اما بیشتر مبتنی بر داده‌های با کلاس معین اسـت و مـی تـوان آن را بیشـتر روشـی نظارتی تا بدون نظارت در نظر گرفت.

در مقاله Lughofer, Weigl et al. (2016) روشی با ترکیب دو تکنیک نیمه نظارتی و بدون نظارت برای شناسایی رانش مفهوم ارائه شده است. روش اول که یک روش نیمه نظارتی است با استفاده از فیلتر های یادگیری فعال داده هـایی کـه بـرای کـلاس‌بنـدی نظـارتی مناسب‌تر هستند را شناسایی می‌کند.این داده ها، داده هایی هستند که کلاس‌بندی‌کننده کمترین میزان اطمینان را بـرای کـلاس‌بنـدی آنها دارد و معمولا داده های نزدیک مرزهای کلاس‌بندی هستند.روش دوم براساس همپوشانی توزیع قطعیت دو کلاس مختلف پیش‌بینی کلاس‌بندی کننده کار می‌کند. در هر دو روش برای شناسایی رانش از آزمون PHT آنلاین و بهبود یافته (Page–Hinkley test) استفاده میشود،در روش اول براساس معنی دار ترین داده ها خطای پیش‌بینی کلاس‌بندی کننده با این آزمون تخمین زده مـی‌شـود.در روش دوم فرض بر این است که توزیع قطعیت کلاس‌های مختلف جدا از هم هستند و به صورت توزیع های گوسی مجـزا فـرض مـی‌شـوند،تغییرات همپوشانی این کلاس‌ها به عنوان متغیر مورد بررسی رانش در آزمون PHT استفاده می گردد.با اینکه روش بدون نظـارت تنهـا قـادر بـه شناسایی رانش‌های درون فضای ورودی است،با استفاده در کنار روش نیمه نظارتی مطرح شده، تفکیک رانش‌های فضای ورودی و رانش-های هدف از هم را فراهم می‌کند.

در Kim and Park (2017) نیز روشی برای شناسایی رانش مفهوم با تعداد محدود نمونه با کلاس حقیقی و بـدون کـلاس حقیقـی ارائه کرده است.این روش براساس تخمین احتمالی نتیجه کلاس‌بندی رانش مفهوم را شناسایی می‌کند.در این روش نمونه هـای تصـادفی انتخاب شده و بردار اطمینان کلاس‌بندی‌کننده محاسبه می‌شود،با استفاده از پنجره،تغییرات اطمینان بررسـی مـی‌شـود و از ایـن طریـق

رخداد رانش شناسایی می‌شود.برای حالت بدون نمونه باکلاس معین،نگارنده پیشنهاد می‌دهد که از میانگین کی(k-means) بـرای ایجـاد ساختار خوشه ای استفاده شود تا به کلاس‌بندی‌کننده مجازی دست پیدا کنیم.ادامه روند مشابه حالت نیمه نظارتی مطرح شده است.

در Yu, Wang et al. (2018) روشی نیمه نظارتی بر اساس چنـد آزمـون فرضـیه سـری ارائـه شـده اسـت.در ایـن مقالـه دو روش Hierarchical Hypothesis Testing with Classification Uncertainty (HHT-CU) و Hierarchical Hypothesis Testing with Attribute-wise "Goodness-of-fit" (HHT-AG) بـرای شناسـایی رانـش مفهـوم پیشـنهاد شـده اسـت.HHT-CU در لایـه تشخیص از نابرابری هافدینگ که تغیرات عدم قطعیت کلاس‌بندی‌کننده را بررسی می‌کند، استفاده می‌کند و بـرای لایـه تصـدیق هـم از آزمون جایگشتی برای تایید رخداد رانش،استفاده می‌شود.HHT-AG در لایه تشخیص از آزمون Kolmogorov-Smirnov روی توزیـع هر یک از ویژگی‌ها رخداد رانش را بررسی می‌کند. در لایه تصدیق کلاس حقیقی داده‌ها برای رد یا تایید رخداد رانش درخواست می‌شود.

در Hosseini, Gholipour et al. (2016) یک کلاس‌بندی‌کنند نیمه نظارتی دسته‌جمعی برای شناسایی رانـش مفهـوم ارائـه شـده است.در این روش،هر کلاس‌بندی‌کننده یک مفهوم منفرد را در خود دارد.وزن کلاس‌بندی‌کننده‌ها بـر اسـاس بسـته داده‌هـایی کـه فقـط تعدادی از آنها کلاس حقیقی دارند تعیین شده و از همین داده‌ها برای آپدیت کردن کلاس‌بندی‌کننده‌ها استفاده می‌شود.

روش Haque, Khan et al. (2015) بـا نـام SAND (Semi-supervised Adaptive Novel Class Detection and Classification over Data Stream) از دو ماژول شناسایی رانـش‌مفهـوم و شناسـایی داده پـرت(Outlier detection) تشـکیل شـده است،این روش هم توانایی شناسایی رانش مفهوم را داشته و هم می‌تواند با جمع‌آوری داده هایی که در مدل کـلاس‌بنـدی‌کننـده دسـته-جمعی کلاس آن‌ها تعیین نمی‌شود؛کلاس جدید موجود را بـه مجموعـه اضـافه کنـد.این روش براسـاس کـا نزدیـک تـرین (k-nearest neighbor) کلاس‌بندی‌کننده دسته‌جمعی،کلاس نمونه ها را پیش‌بینی کرده و میزان اطمینان هر کلاس‌بندی کننده را تخمین میزنـد.با کنترل تغییرات اطمینان،این روش قادر است رانش مفهوم را شناسایی کند. بعد از شناسایی رانش SAND فقط کلاس حقیقی داده هایی را درخواست می‌کند که اطمینان خیلی پایینی دارند و باقی داده‌ها براساس همان مدل پیش‌بینی کلاس‌بندی مـی‌شـوند.با وجـود اینکـه روش SAND از تعداد نمونه با کلاس حقیقی کمتری استفاده می‌کند و درستی پیش‌بینی خوبی هم دارد،اما از نظر هزینـه محاسـباتی و زمان محاسبه به علت حجم بالای محاسبات کاملا کـاربردی نبـوده اسـت.از همـین رو،روش SAND در Haque, Khan et al. (2016) بهبود پیدا کرده است.روش جدید که بـا نـام Efficient Concept Drift and Concept Evolution Handling over Stream Data (ECHO) شناخته می‌شود.این روش عملکردی مانند روش SAND دارد با این تفاوت که ماژول شناسایی رانش مفهوم حال دیگر در هر بار محاسبه میزان آستانه اطمینان مدل دسته‌جمعی فعال نشده و فراخوانی آن وابسته به میزان اطمینان مجموعه تست خواهد بود.

۳,۴-روش‌های بدون نظارت

روش های بدون نظارت شناسایی رانش مفهوم مناسب شرایطی هستند که بازخورد نتایج پیش‌بینی کلاس‌های داده بـا تـاخیر همـراه است.این روش ها با بررسی تغییرات ابعاد داده رانش‌مفهوم را شناسایی می‌کنند.

در مقاله Liu, Lu et al. (2018) روشی برای شناسایی رانش های منطقه‌ای به صورت بدون نظـارت مطـرح شـده اسـت.طبق گفتـه نگارنده، بیشتر روش های شناسایی رانش مفهوم در تلاش برای شناسایی رانش های عمومی هستند و رانش هایی که بـه علـت تغییـرات چگـالی منطقـه ای بـه وجـود مـی‌آینـد را نمـی‌تواننـد شناسـایی کننـد.روش Nearest Neighbor-based Density Variation Identification (NN-DVI) یک روش مبتنی بر تخمین چگالی منطقه ای است و از سه ماژول جزبندی فضا بر اساس کا نزدیـک تـرین همسایه(kNN-based space partitioning)،تابع فاصله میزان اختلاف چگالی و آزمون معنی‌داری آماری تشکیل شده است.این روش بـا مقایسه بین پنجره داده های جدید و پنجره داده های قدیمی تر کار مـی‌کنـد.در مـاژول اول بـر اسـاس همسـایگی هـر دادهx،آن داده و همسایه های آن تشکیل یک ذره نمونه(instance particle)را میدهند. بعد گروه های ذره نمونه بر اساس همسایگی مشترک ایجاد شـده و در نتیجه هر داده با مجموعه ای از ذره نمونه های دارای اشتراک نمایش داده میشود. این باعث مـی‌شـود کـه هـر ذره نمونـه مشـترک جزئیات بیشتری از نمونه را معرفی کرده و تغییرات جزیی نیز بهتر مشاهده شوند.در ماژول دوم تفاوت میان داده هـا بـر اسـاس مجمـوع

تفاوت تعداد ذره نمونه آنها محاسبه میشود.در ماژول سوم آزمونی برای تایید رانش انجام میگیرد.نویسنده مقاله فرض می‌کند کـه مقـادیر تفاوت های نمونه های مورد مقایسه در یک توزیع نرمال قرار می‌گیرند؛ از این رو از برآورد درست‌نمایی بیشینه(max likelihood)بـرای تخمین میانگین و واریانس توزیع تفاوت نمونه های مورد بررسی استفاده می‌کند. سپس تابع توزیع انباشـته(cumulative distribution function)با مقادیر آستانه تعریف شده توسط کاربر مورد بررسی قرار میگیرد و رخداد رانش مفهوم تایید می‌گردد.

در Bashir, Petrovski et al. (2017) روشـی دو مرحلـه‌ای بـا نـام Unsupervised Change Detection for Activity Recognition برای شناسایی رانش مفهوم به صورت بدون نظارت ارائه شده است.این روش مبتنی بر این فرض است که رانش مفهوم به شکل تفاوت زیاد بین داده‌های جدید منسوب به یک کلاس با نمونه های آموزشی آن کلاس،ظاهر می‌شود.در مرحله اول ایـن روش یـک کلاس‌بندی کننده تمرین میدهد و ویژگی هر یک از کلاس ها را تعیین می‌کند.در مرحله دوم داده ها به صورت بسته ای دریافت شـده و بررسی می‌شوند تا تفاوت آنها با داده های پیشین تعیین شود.با اینکه این روش در این مقاله برای شناسایی رفتار فیزیکی نظیـر دویـدن و نشستن و... مطرح شده است اما تاکیدی بر انحصار آن برای شناسایی حرکت بیان نشده است.

در مقاله Sethi and Kantardzic (2015)روشی با نـام Margin Density Drift Detection مطـرح شـد کـه بـر اسـاس کنتـرل تغییرات در مرز اختلاف کلاس‌بندی‌کننده(classifier's margin)رانش مفهوم را شناسایی می‌کند.مرز اخـتلاف بـه ایـن صـورت تعریـف میشود: ناحیه ای است در نزدیکی به کرانه های مدل و پیش‌بینی ها برای این ناحیه ها قطعیت بسیار پایینی دارد.فرض بر این اسـت کـه تغییرات چگالی این ناحیه ها بیشتر از آستانه تعریف شده نشان از رخداد رانش است.با استفاده از تعداد نمونه هایی که در این ناحیـه هـا قرار دارند چگالی مرز اختلاف محاسبه می‌شود.با استفاده از دیتاست اولیه و روش SVM کلاس‌بندی‌کننده اولیه تولید شده و با توجه بـه اینکه از روش SVM استفاده می‌شود، مرز های اختلاف طبقه‌بندی‌کننده مشخص خواهد شد.در مرحله بعد بیشترین و کمتـرین مقـادیر چگالی بعد از آخرین شناسایی رانش مفهوم محاسبه شده و تحت نظارت قرار میگیرد و برای هر دسته داده جدید که به سیستم می‌رسـد چگالی آنها محاسبه شده و با مقایسه آن با میزان آستانه تعریف شده می‌توان رخداد رانش را شناسایی کرد.بعـد از اینکـه رخـداد رانـش شناسایی شد مدل بر اساس نمونه هایی که کلاس حقیقی آنها مشخص هستند آپدیت میشود. در این مرحله به کلاس حقیقـی داده هـا نیازمنـد هسـتیم.یکی از مشـکلات اصـلی ایـن روش نیازمنـدی بـه روش یـادگیری SVM اسـت.در سـال 2017 در مقالـه Sethi and Kantardzic (2017) روش پیشنهادی برای اینکه مشکلات پیشین را نداشته باشد،بهبود یافت.روش جدیـد بـا نـام MD3 Ensemble Generic Margin (MD3-EGM) شناخته میشود. در این روش با استفاده از کلاس‌بندی‌کننـده دسـته‌جمعـی مـرز اخـتلاف عمـومی شناسایی می‌شود که مجموعه ای از نمونه های ناشناخته است که بیشترین عدم قطعیت را برای پیش‌بینـی دارنـد.این عـدم قطعیـت، بـر اساس میزان اختلاف کلاس‌بندی‌کننده های مدل دسته‌جمعی برای پیش‌بینی کلاس یک داده بدست می‌آید.شناسایی رانش نیز بر اساس همین افزایش میزان اختلاف برای کلاس‌بندی یک داده ناشناخته صورت می‌گیرد.از آنجایی که نرخ هشدار اشتباه این روش می‌تواند زیاد باشد؛ یک ماژول تایید رانش نظارتی به این روش اضافه میشود و آن را به روشی نیمه نظارتی تبدیل می‌کند.

در Costa, Albuquerque et al. (2018) روشی برای شناسایی رانش مفهوم با فرض اینکه تغییر چگـالی معنـی دار تـرین داده هـا می‌تواند نشانه رانش مفهوم باشد،مطرح شد.در این روش بر اساس اطمینان کلاس‌بندی‌کننده برای داده هـا حـدود مجـازی تعیـین مـی-شود.نمونه‌هایی که اطمینان آن‌ها از میزان آستانه کمتر باشد طبق یادگیری فعال(Active Learning) برای محاسبه چگالی انتخاب می-شوند.میزان چگالی داده های جدید با میزان های بیشینه و کمینه موجود چگالی مقایسه شده و در صورتی که از میزان آستانه تایید شده بیشتر باشد؛رخداد رانش مفهوم شناسایی می‌شود.

در dos Reis, Flach et al. (2016) روشـی بـدون نظـارت بـا بکـارگیری آزمـون کولموگـوروف-اسـمیرنوف(Incremental Kolmogorov–Smirnov)مطـرح شـد.ایده اصـلی ایـن روش،اعمـال ایـن آزمـون بـرای هریـک از ویژگـی‌هـای داده بـه صـورت مجـزا است.شناسایی تغییر در یک ویژگی ممکن است که برای شناسایی رانش مفهوم کافی باشد.از ضعف های ایـن روش ایـن اسـت کـه قـادر نخواهد بود رانش های مفهومی را شناسایی کند که نیازمند بررسی ویژگی‌های نمونه ها درکنار هم هست.

در Mustafa, Ayoade et al. (2017) یک روش برای شناسایی کلاس جدید در داده با نام Nonparametric Multidimensional Drift Detection Method (NM-DDM) ارائه شد که از denoising autoencoders که یکی از الگوریتم هـای شـبکه هـای عصـبی است؛ استفاده میکند.در این مقاله یک روش برای شناسایی نقطه تغییر و رانش مفهـوم ارائـه شـد کـه در اینجـا مـورد نظـر ماسـت.روش شناسایی رانش مفهوم بر اساس محاسبه نسبت درست‌نمایی (log-likelihood ratio)بین دو پنجره،قبل از نقطه احتمالی رانش و بعـد از آن، کار میکند.نسبت آزمون‌درست نمایی برای هریک از ویژگی های داده(ابعاد) تخمین زده شده و در صورتی که بیشترین نسبت محاسبه شده میان دو پنجره،بیشتر از مقدار آستانه باشد،رخداد رانش تایید شده و داده های پنجره جدیدتر بـرای یـادگیری مـورد اسـتفاده قـرار میگیرد.برای مرحله آپدیت مدل تنها کلاس واقعی داده هایی مورد نیاز است که میزان اطمینان آنها در کلاس‌بندی‌کننده پایین بوده است و باقی داده ها با همان کلاس پیش‌بینی شده مورد بررسی قرار میگیرند.

Plover روشی اسـت کـه در de Mello, Vaz et al. (2019) مطـرح مـی‌شـود کـه بـر اسـاس معیـارهـای آمـاری مختلـف نظیـر میانگین،واریانس، .... و طیف توانی (power spectrum) رانش‌مفهوم را شناسایی می‌کند.نگارنده در این مقاله تحلیلی بر مبـانی نظـری و تضمین های یادگیری در شناسایی رانش مفهوم بدون نظارت انجام داده است.آن‌ها نتیجه‌گیری کرده اند؛درحالی که روش‌های نظارتی بـا نظریه‌های یادگیری آماری پشتیبانی می‌شوند؛روش های بدون نظارت وابسته به معیار هایی درونی و بیرونی هستند که چنین پشتوانه ای ندارند.معیار های درونی نظیر فشردگی خوشه و فاصله گرانیگاه که از خود ساختار داده قابل محاسبه بوده،هسـتند.در مقابـل معیـار هـای بیرونی وابسته به کلاس حقیقی داده است که در بسیاری از شرایط برای روش‌های بدون نظارت ممکن نخواهد بود.

در Pinagé, dos Santos et al. (2020) یکی دیگر از روش های بدون نظارت معرفی شد.DSDDیـا Dynamic Selection Drift Detector روشی است که به صورت پویا از مجموعه کلاس‌بندی‌کننده های دسته جمعی،کلاس‌بندی‌کننده مناسب هـر داده را انتخـاب کرده و با کنترل شبه‌خطا(Pseudo Error)می تواند رانش مفهوم را شناسایی کند و از سه ماژول مختلف تشکیل شده اسـت.ماژول هـای ساخت مدل دست جمعی(Ensemble Creation)،انتخاب پویا(Dynamic Selection)و شناسایی(Detection).در ماژول اول با استفاده از دیتاست اولیه با روش Online Bagging بهبود یافته یک مدل دسته‌جمعی از کلاس‌بندی‌کننده ها ایجاد می‌شود و فرض بر این است که هرکدام از این کلاس‌بندی کننده ها در یک ناحیه از داده ها بهترین عملکرد را دارد.ماژول دوم بدین صورت کار می‌کند که وقتی یـک داده جدید و بدون کلاس معلوم به ماژول وارد میشود، براساس اینکه کدام کلاس‌بندی‌کننده بهترین عملکرد را برای داده هـای همسـایه آن داده در دیتاست اعتبار سنجی بر اساس (k nearest neighbor)داشته،آن کلاس‌بندی کننده برای پیش‌بینی کلاس آن داده انتخـاب میشود.کلاسی که کلاس بندی‌کننده منتخب در ماژول دوم برای داده پیش‌بینی می‌کند بـه عنـوان کـلاس شـبه درسـت آن داده تاییـد میشود(pseudo true).در ماژول بعدی برای هریک از کلاس‌بندی کننده های مدل دسته جمعی یکـی از روش هـای تشـخیص رانـش بـا نظارت مانند DDM یا EDDM مورد استفاده قرار می‌گیرد.با اینکه این روش ها برای کارکرد نیازمند کلاس حقیقی داده هسـتند امـا در اینجا با توجه به کلاس شبه درست داده بررسی ها صورت می‌گیرد و در صورتی که تعداد مشخصی ازکلاس‌بندی کننده ها تشخیص رانش دهند؛ رخداد رانش مفهوم تایید شده و دیتاست اعتبارسنجی با توجه به داده های جدید آپدیت شده و مقادیر نرخ خطا و انحـراف معیـار مورد استفاده در DDM ریست می‌شود.در این مقاله از DDM وEDDM استفاده شد اما می توان از هرکدام از روش های مبتنی بر نـرخ خطا برای ماژول تشخیص رانش استفاده کرد.

با توجه به اهمیت کارکرد این روش‌ها در حالتی که کلاس حقیقی داده ها با تاخیر همراه است؛بسیاری از روش‌هـا از تکنیـک هـای مختلفی در کنار هم استفاده می‌کنند که فاصله گذاری روش‌ها از یکدیگر را سخت می‌کند.در این قسمت سعی شد تا با توجـه بـه میـزان وابستگی به کلاس حقیقی داده ها،روش ها به این سه گروه تقسیم شوند.

۴-نتیجه گیری

در این گزارش سعی شد دسته بندی نظارتی،بدون نظارت و نیمه نظـارتی کـه در مقالـه Khamassi, Sayed-Mouchaweh et al. (2018) مطرح شده بود با بررسی روش‌های مطرح شده بعد از سال ۲۰۱۵ به روز شود. این تقسیم‌بندی،با توجه به اندازه نیاز این روش-

ها به کلاس حقیقی داده ها برای کار کرد،انجام شده است. در بیشتر موقعیت‌های دنیای واقعی که کلاس حقیقی داده ها یـا در دسـترس نیست یا با تاخیر در دسترس است،روش های نیمه نظارتی و بدون نظارت توانایی بیشتری برای شناسـایی رانـش‌مفهـوم دارنـد.با بررسـی روش‌های نظارتی،می توان دید که این روش ها توانایی شناسایی رانش مفهوم با دقتی بالا را داشته و می توان بیشـتر گونـه‌هـای رخـداد رانش‌مفهوم را با آن‌ها تشخیص داد،در مقابل فراهم کردن کلاس حقیقی همه داده ها هزینه سنگینی است که این روش ها برای رسـیدن به این دقت باید بپردازند.روش های بدون نظـارت و نیمـه نظارتی،بـا کـار کـردن در شـرایطی کـه کـلاس حقیقـی داده هـا در دسـترس نیست،امکان شناسایی رانش مفهوم در شرایط دنیای واقعی را فراهم می‌کنند اما در مقابل توانایی شناسایی رانش مفهوم آن ها پـایین تـر بوده و هرکدام توانایی شناسایی دامنه محدودی از انواع شکل‌های رخداد رانش‌مفهوم را دارند.بر همین اساس،توازنی بین عملکرد و هزینه بین روش‌های مختلف شناسایی رانش‌مفهوم مشاهده می‌شود که انتخاب بهترین روش را مشکل خواهد کرد. از دیگر مشکلات می توان به نبود معیار و مقیاس مطرحی برای انتخاب دیتاست و بررسی عملکرد روش‌های شناسایی رانش‌مفهوم در این دیتاست ها اشاره کرد.با ایجاد این چارچوب، بررسی عملکرد روش‌ها آسان تر خواهد بود.معیار های بررسی هر روش مانند دقت کلاس‌بندی به تنهایی کافی نیستند.نیاز به معیارهای بیشتر برای سنجش میزان عملکرد روش‌های شناسایی از دیگر مشکلات پیرامون ایـن بحـث اسـت.در آخر،معرفـی و ابـداع روش‌هایی که توانایی شناسایی رانش مفهوم با توجه به نا هنجاری های داده نظیر وابستگی موقت(temporal dependence)،عدم تعـادل کلاس(Class imbalance) و جریان های داده با مقادیر از دست رفته(missing value) به شدت می تواند به شناسایی رانـش مفهـوم و مطالعه بهتر داده ها کمک کند.

روش های شناسایی رانش که در مدل دسته‌جمعی استفاده می‌شود،تکنیکی است که شانس شناسایی انواع رانش مفهوم را بیشتر می کند به خصوص برای روش های بدون نظارت که ضعف آن‌ها در شناسایی انواع گونه‌های رانش مفهوم است.بررسی مجزا این روش‌ها خود مـی تواند موضوعی برای مقاله ای مروری در حوزه شناسایی رانش مفهوم باشد.

۵-مراجع

[1]Khamassi, I., M. Sayed-Mouchaweh, M. Hammami and K. Ghédira (2018). "Discussion and review on evolving data streams and concept drift adapting." Evolving systems **9**(1): 1-23.
[2]Lu, J., A. Liu, F. Dong, F. Gu, J. Gama and G. Zhang (2018). "Learning under concept drift: A review." IEEE Transactions on Knowledge and Data Engineering **31**(12): 2346-2363.
[3]Hu, H., M. Kantardzic and T. S. Sethi (2020). "No Free Lunch Theorem for concept drift detection in streaming data classification: A review." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(2): e1327.
[4]Wares, S., J. Isaacs and E. Elyan (2019). "Data stream mining: methods and challenges for handling concept drift." SN Applied Sciences **1**(11): 1412.
[5]Gemaque, R. N., A. F. J. Costa, R. Giusti and E. M. Dos Santos (2020). "An overview of unsupervised drift detection methods." Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **10**(6): e1381.
[6]Liu, A., G. Zhang and J. Lu (2017). Fuzzy time windowing for gradual concept drift adaptation. 2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE), IEEE.
[7]Xu, S. and J. Wang (2017). "Dynamic extreme learning machine for data stream classification." Neurocomputing **238**: 433-449.
[8]Gama, J., P. Medas, G. Castillo and P. Rodrigues (2004). Learning with drift detection. Brazilian symposium on artificial intelligence, Springer.
[9]Wang, H. and Z. Abraham (2015). Concept drift detection for streaming data. 2015 International Joint Conference on Neural Networks (IJCNN) ,IEEE.
[10]Zhang, Y., G. Chu, P. Li, X. Hu and X. Wu (2017). "Three-layer concept drifting detection in text data streams." Neurocomputing **260**: 393-403.
[11]Alippi, C., G. Boracchi and M. Roveri (2016). "Hierarchical change-detection tests." IEEE transactions on neural networks and learning systems **28**(2): 246-258.
[12]Yu, S. and Z. Abraham (2017). Concept drift detection with hierarchical hypothesis testing. Proceedings of the 2017 SIAM International Conference on Data Mining, SIAM.
[13]de Lima Cabral, D .R. and R. S. M. de Barros (2018). "Concept drift detection based on Fisher's Exact test." Information Sciences **442**: 220-234.
[14]Nishida, K. and K. Yamauchi (2007). Detecting concept drift using statistical testing. International conference on discovery science, Springer.

[15]Pesaranghader, A., H. L. Viktor and E. Paquet (2018). McDiarmid drift detection methods for evolving data streams. 2018 International Joint Conference on Neural Networks (IJCNN), IEEE.
[16]Pesaranghader, A. and H. L. Viktor (2016). Fast hoeffding drift detection method for evolving data streams. Joint European conference on machine learning and knowledge discovery in databases, Springer.
[17]Wang, X., Q. Kang, M. Zhou and S. Yao (2018). A multiscale concept drift detection method for learning from data streams. 2018 IEEE 14th International Conference on Automation Science and Engineering (CASE), IEEE.
[18]Dehghan, M., H. Beigy and P. ZareMoodi (2016). "A novel concept drift detection method in data streams using ensemble classifiers." Intelligent Data Analysis **20**(6): 1329-1350.
[19]Sidhu, P. and M. Bhatia (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.
[20]Khamassi, I., M. Sayed-Mouchaweh, M. Hammami and K. Ghédira (2019). A new combination of diversity techniques in ensemble classifiers for handling complex concept drift. Learning from data streams in evolving environments, Springer**:** 39-.61
[21]Khamassi, I., M. Sayed-Mouchaweh, M. Hammami and K. Ghédira (2015). "Self-adaptive windowing approach for handling complex concept drift." Cognitive Computation **7**(6): 772-790.
[22]Mahdi, O. A., E. Pardede and J. Cao (2018). Combination of information entropy and ensemble classification for detecting concept drift in data stream. Proceedings of the Australasian Computer Science Week Multiconference.
[23]Krawczyk, B. and A. Cano (2018). "Online ensemble learning with abstaining classifiers for drifting and noisy data streams." Applied Soft Computing **68**: 677-692.
[24]Jaworski, M., P. Duda and L. Rutkowski (2017). On applying the restricted Boltzmann machine to active concept drift detection. 2017 IEEE Symposium Series on Computational Intelligence (SSCI), IEEE.
[25]Lobo, J. L., J. Del Ser, I. Laña, M. N. Bilbao and N. Kasabov (2018). Drift detection over non-stationary data streams using evolving spiking neural networks. International symposium on intelligent and distributed computing, Springer.
[26]Astudillo, C. A., J. I. González, B. J. Oommen and A. Yazidi (2016). Concept drift detection using online histogram-based bayesian classifiers. Australasian Joint Conference on Artificial Intelligence, Springer.
[27]Ren, S., B. Liao, W. Zhu and K. Li (201" .(8Knowledge-maximized ensemble algorithm for different types of concept drift." Information Sciences **430**: 261-281.
[28]Lughofer, E., E. Weigl, W. Heidl, C. Eitzinger and T. Radauer (2016). "Recognizing input space and target concept drifts in data streams with scarcely labeled and unlabelled instances." Information Sciences **355**: 127-151.
[29]Kim, Y. and C. H. Park (2017). "An efficient concept drift detection method for streaming data under limited labeling." IEICE Transactions on Information and systems **100**(10): 2537-2546.
[30]Yu, S., X. Wang and J. C. Principe (2018). "Request-and-reverify: Hierarchical hypothesis testing for concept drift detection with expensive labels." arXiv preprint arXiv:1806.10131.
[31]Hosseini, M. J., A. Gholipour and H. Beigy (2016). "An ensemble of cluster-based classifiers for semi-supervised classification of non-stationary data streams." Knowledge and information systems **46**(3): 567-597.
[32]Haque, A., L. Khan and M. Baron (2015). Semi supervised adaptive framework for classifying evolving data stream. Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer.
[33]Haque, A., L. Khan, M. Baron, B. Thuraisingham and C. Aggarwal (2016). Efficient handling of concept drift and concept evolution over stream data 2016 .IEEE 32nd International Conference on Data Engineering (ICDE), IEEE.
[34]Liu, A., J. Lu, F. Liu and G. Zhang (2018). "Accumulating regional density dissimilarity for concept drift detection in data streams." Pattern Recognition **76**: 256-272.
[35]Bashir, S. A., A. Petrovski and D. Doolan (2017). "A framework for unsupervised change detection in activity recognition." International Journal of Pervasive Computing and Communications.
[36]Sethi, T. S. and M. M. Kantardzic (2015). Don't Pay for Validation :Detecting Drifts from Unlabeled data Using Margin Density. INNS Conference on Big Data.
[37]Sethi, T. S. and M. Kantardzic (2017). "On the reliable detection of concept drift from streaming unlabeled data." Expert Systems with Applications **82**: 77-99.
[38]Costa, A. F. J., R. A. Albuquerque, n. Saraiva and E. M. dos Santos (2018). A drift detection method based on active learning. 2018 International Joint Conference on Neural Networks (IJCNN), IEEE.
[39]dos Reis, D. M., P. Flach, S. Matwin and G. Batista .(2016)Fast unsupervised online drift detection using incremental kolmogorov-smirnov test. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.
[40]Mustafa, A. M., G. Ayoade, K. Al-Naami, L. Khan, K. W. Hamlen, B. Thuraisingham and F. Araujo (2017). Unsupervised deep embedding for novel class detection over data stream. 2017 IEEE International Conference on Big Data (Big Data), IEEE.
[41]de Mello, R. F., Y. Vaz, C. H. Grossi and A. Bifet (2019). "On learning guarantees to unsupervised concept drift detection on data streams." Expert Systems with Applications **117**: 90-102.
[42]Pinagé, F., E. M. dos Santos and J. Gama (2020). "A drift detection method based on dynamic classifier selection." Data Mining and Knowledge Discovery **34**(1): 50-74.