ارائه راهکار بهبود یافته مبتنی بر رانش مفهوم به منظور به روزرسانی پروفایل مشتریان در سیستم‌های کشف تقلب

فاطمه کردی اردستانی[۱]، سامان سیادتی[۲]، حسین هاشم زاده[۳]

چکیده

یک سیستم کشف تقلب مرسوم از ابزارهای اتوماتیک و فرایندهای انسانی تشکیل شده است که ابزارهای اتوماتیک بر پایه قوانین کشف تقلب هستند. این ابزارها کل تراکنش‌های ورودی را تحلیل می‌کنند و بر چسب جعلی یا واقعی به آنها اختصاص می‌دهند. فرایندهای انسانی توسط بازرسان ایجاد می‌شود به این معنی که تراکنش‌ها با امتیاز جعلی بالا را بررسی می‌کنند و یک بازخورد برای آنها در نظر می‌گیرند.[1] در سال‌های اخیر با هدف امن نمودن محیط تجارت الکترونیکی، تحقیقات بسیاری در حوزه کشف تقلب انجام و روش‌های مختلفی ارائه شده است. یکی از چالش‌های مطرح در این روش‌ها عدم وفق پذیری آن‌ها با تغییر رفتار مشتریان و متقلبان می‌باشد. این تغییرات تحت عنوان رانش مفهوم شناخته می‌شوند.
بطور کلی رانش مفهوم اشاره به چالش تغییر توزیع داده‌ها در طول زمان اشاره دارد و باعث می‌شود استفاده از داده‌های قدیمی برای تعیین وضعیت داده‌های جدید مناسب نباشد. برای مدیریت رانش مفهوم دراین روش‌ها دو تأثیر متناقض وجود دارد؛ از طرفی در حضور رانش مفهوم احتمال منسوخ شدن داده‌های قدیمی زیاد است و استفاده از آن‌ها افت دقت را بدنبال دارد و از طرف دیگر تاریخچه رفتاری مشتری و سازماندهی مناسب آن در ساخت پروفایل آن‌ها نقش اساسی دارد. تغییرات در رفتار مشتری ممکن است منجر به هشدارهای نادرست در تکنیک‌های کشف تقلب شود. در حقیقت رانش مفهوم منجر به افت شدید در دقت سیستم‌های کشف تقلب خواهد شد.[2]

در این تحقیق قصد داریم روشی را برای به روزرسانی پروفایل مشتریان با در نظر گرفتن رانش مفهوم طراحی کنیم. از روش‌های سازگاری رانش مفهوم می‌توان به روش پنجره کشویی اشاره کرد. پس از مدلسازی، با استفاده از الگوریتم‌های فراابتکاری، آموزش طبقه بندها بر اساس جریان داده‌های اخیر و تاخیری انجام می‌گردد. در این پژوهش با ارائه روشی بهبود یافته مبتنی بر پنجره کشویی افزایش دقت در سیستم‌های کشف تقلب نشان داده می‌شود.

کلمات کلیدی: کشف تقلب، پروفایل، رانش مفهوم[۴]، داده کاوی، پنجره کشویی[۵]، الگوریتم فراابتکاری[۶]

[۱] دانشجویی کارشناسی ارشد مدیریت فناوری اطلاعات، دانشگاه آزاد اسلامی واحد تهران جنوب/ تحلیلگر ارشد هوش تجاری شرکت تأمین خدمات سیستم‌های کاربردی کاسپین، ardestany@gmail.com
[۲] استادیار دانشگاه آزاد اسلامی واحد تهران جنوب، dr.samansiadati@gmail.com
[۳] کارشناسی ارشد مهندسی برق، دانشگاه صنعتی شریف، h. hashemzade@gmail.com

[۴] Concept Drift
[۵] Sliding window
[۶] Metaheuristic Algorithm

مقدمه

توسعه فناوری‌های نوین اطلاعاتی و ارتباطی و وابستگی روز افزون بانک‌ها و نظام‌های پرداخت به این فناوری‌ها، در کنار تسهیل فرایندهای بانکی و مدیریتی، بستر ارتکاب جرائم سایبری و سوءاستفاده از این خدمات را نیز فراهم می‌سازد. کارت اعتباری تراکنش‌های آنلاین را ساده‌تر و راحت‌تر ساخته است. با این حال، روند رو به رشد تقلب در تراکنش‌ها که منجر به زیان‌های مالی بسیار در هر سال می‌شود، وجود دارد. پیش بینی شده است که نرخ رشد زیان‌های مالی سالانه به عدد دو رقمی تا سال ۲۰۲۰ افزایش خواهد یافت [3]. شرکت‌های پردازش تراکنش‌های الکترونیکی بایستی هر گونه رفتار جعلی را به منظور حفظ اعتماد مشتریان و ایمنی کسب و کار خود تشخیص دهند.

عدم وفق پذیری سیستم‌های تشخیص تقلب با تغییرات رفتاری بهنجار مشتریان، منجر به تولید تعداد بالای هشدارهای نادرست می‌شود ؛ بدین معنا که تراکنش‌های عادی زیادی به اشتباه تقلب تشخیص داده می‌شوند که این امر برای دارندگان حساب آزار دهنده است. این تغییرات تحت عنوان رانش مفهوم شناخته می‌شوند و در دسته‌ای از راهکارها با نام راهکارهای کاوش جریان داده ۱در حوزهای کاربردی متفاوت مورد توجه محققین قرار گرفته است. در این راهکارها جریان‌های داده‌ای ترتیبی از نمونه‌ها هستند که به طور پیوسته و با گذر زمان وارد می‌شوند و توزیع داده‌های آن‌ها ممکن است در طول زمان دست خوش تغییر شود.[2] با توجه به تغییر توزیع داده‌ها در طول زمان، داده‌های مربوط به گذشته ممکن است غیر مرتبط شوند و برای خلاصه سازی های جاری مناسب نباشند. این امر نیاز به ارائه راهکارهای مدیریت رانش مفهوم را روشن می‌سازد. در حقیقت رانش مفهوم منجر به افت شدید در دقت سیستم‌های کشف تقلب خواهد شد و در این مقاله سعی می‌شود دقت هشدار که نگرانی اصلی محققان است، با روش پیشنهادی به‌صورت قابل ملاحظه‌ای بهبود یابد.

ادبیات موضوع

طبق تحقیقات [8] که در آنها مروری بر روش‌های بکار رفته برای تشخیص تقلب صورت گرفته است این روش‌ها به دو دسته کلی تقسیم می‌شوند:

۱- روش‌های مبتنی بر رفتار نادرست

در این روش که شناسایی سو استفاده نیز گفته می‌شود، تراکنش‌های متقلبانه شناخته شده در یک الگو قرار می‌گیرند و بعد از آن هر تراکنشی مانند آن عمل کند بعنوان تراکنش تقلبی تشخیص داده می‌شود.

۲- روش‌های مبتنی بر رفتار نابهنجار

در این روش که شناسایی نابهنجاری نیز نامیده می‌شود، رفتارهای طبیعی کاربر پس از اولین استفاده در پروفایل او ذخیره می‌شود و از این پس به عنوان معیاری برای تشخیص انحراف فعالیت‌های اخیر کاربر از آن استفاده می‌شود.

---

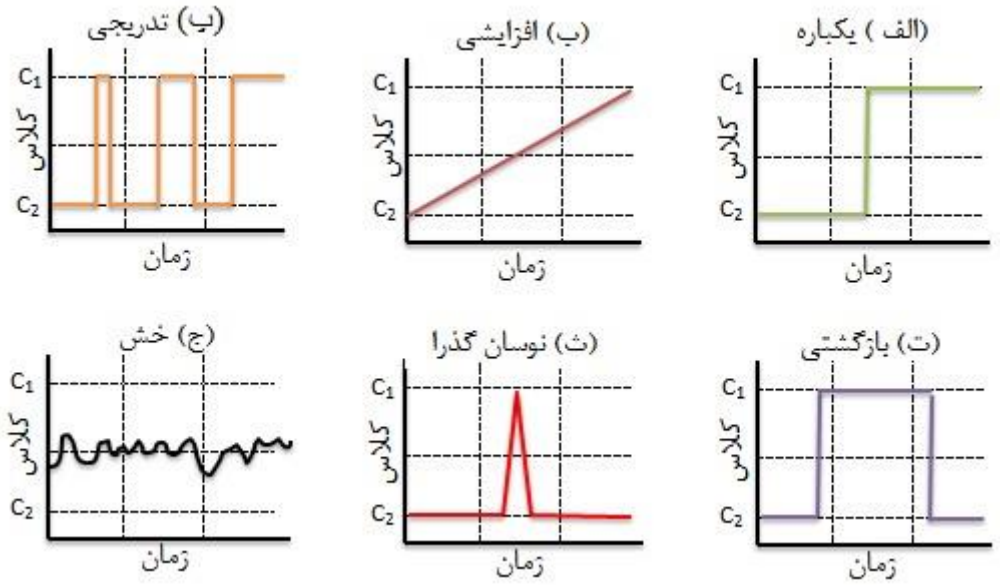
۱ Data Stream mining

تحقیق [5] نمونه‌هایی از روش‌های مبتنی بر کشف سوء استفاده را ارائه داده است. در این دسته از راهکارها، تنها الگوهای شناخته شده تقلب قابل تشخیص هستند. این در حالی است که متقلبان همواره تلاش می‌کنند تا با روش‌های پیچیده و با استفاده از رویه‌های جدیدی در سیستم نفوذ کنند. در مقابل، در دسته دوم که روش‌های مبتنی بر پروفایل نیز نامیده می‌شوند مدل، از روی داده‌های بهنجار ساخته می‌شود و معیار تقلبی بودن یک تراکنش جدید میزان انحراف آن از مدل بهنجار است.
در تحقیقات [6] نمونه‌هایی از روش‌های مبتنی بر کشف نابهنجاری ارائه شده است. در این دسته از راهکارها تغییر در رفتار متقلبان مدیریت می‌شود ولی یکی از چالش‌های همچنان مطرح عدم وفق پذیری این روش‌ها با تغییرات رفتاری بهنجار مشتری می‌باشد. بدین معنا که در این سیستم‌ها هرگونه تغییر در رفتار، الزاماً نشانه تقلب نیست و رفتار دارنده حساب ممکن است به دلایل متفاوتی دچار تغییر شود.[7] برای مثال مبلغ تراکنش‌ها و تعداد تراکنش‌ها به عادت پرداختی دارنده حساب مرتبط می‌باشند که وابسته به درآمد، در دسترس بودن منابع، نحوه زندگی وی و نظایر آن است و معمولاً در طول زمان دچار تغییر می‌شود.

در به‌کارگیری روش‌های مبتنی بر رفتار نابهنجار، ساخت یک پروفایل کارا مهم‌ترین عامل موفقیت است. پروفایل مجموعه‌ای از اطلاعات یا الگوهای مهم و ممتاز است که یک موجودیت را در یک حوزه خاص مشخص می‌کند .[9]در تحقیق [10]دو روش تشخیص مبتنی بر پروفایل و مبتنی بر رفتار نادرست بر روی داده‌های واقعی بکار گرفته‌شده و باهم مقایسه شده‌اند . نتایج تحقیق نشان داده که به دلیل تغییر الگوهای رفتاری متقلبان، روش مبتنی بر رفتار نابهنجار برای تشخیص تقلب در کارت‌های اعتباری مناسب‌تر است. این‌یک تعریف عمومی است که انواع مختلفی از اطلاعات را پوشش می‌دهد. بر اساس نوع اطلاعات مورداستفاده در ساخت پروفایل، می‌توان دودسته کلی برای آن تعریف نمود: ۱. پروفایل عینی ۲. پروفایل رفتاری

پروفایل عینی بر اساس اطلاعات آماری و علاقه‌مندی‌های کاربر ساخته می‌شود، اطلاعاتی از قبیل سن، جنسیت، آدرس، درآمد و... این اطلاعات معمولاً در زمان ثبت‌نام مشتری به دست می‌آید و ممکن است بعدها تغییر کند. اما پروفایل رفتاری شامل الگوها و قوانینی است که مشخصه‌های رفتاری کاربر را توصیف می‌کند. و با تحلیل تراکنش‌های او در طول زمان به دست می‌آید. به‌عنوان‌مثال اینکه خریدهای آخر هفته مشتری معمولاً در بعدازظهر اتفاق می‌افتد، خریدهای مشتری معمولاً از فروشنده‌های محلی انجام می‌شود و چیزهایی از این قبیل. مشخص است که پروفایل رفتاری با تحلیل و بررسی حجم بالایی از تراکنش‌های دارنده کارت به دست می‌آید. گرچه تحلیل پروفایل عینی مشتری می‌تواند در تشخیص تقلب مؤثر باشد اما متقلب معمولاً با الگوهای رفتاری مشتری آشنا نیست و تراکنش‌هایی که با کارت مشتری انجام می‌دهد باعث انحرافاتی از پروفایل رفتاری او خواهد شد. لذا با تحلیل پروفایل رفتاری مشتری می‌توان تراکنش‌های تقلبی را آشکار ساخت و تمرکز این تحقیق نیز بر این نوع پروفایل قرارگرفته است.

از طرفی در جهان واقعی انواع متفاوت رانش مفهوم اتفاق می‌افتد و به علت ویژگی‌های متفاوت، نیاز به استراتژی‌های پیش‌بینی متفاوتی دارند. شکل ۱ که در [2] آمده است انواع تغییر در جریان‌های داده‌ای را نشان می‌دهد.



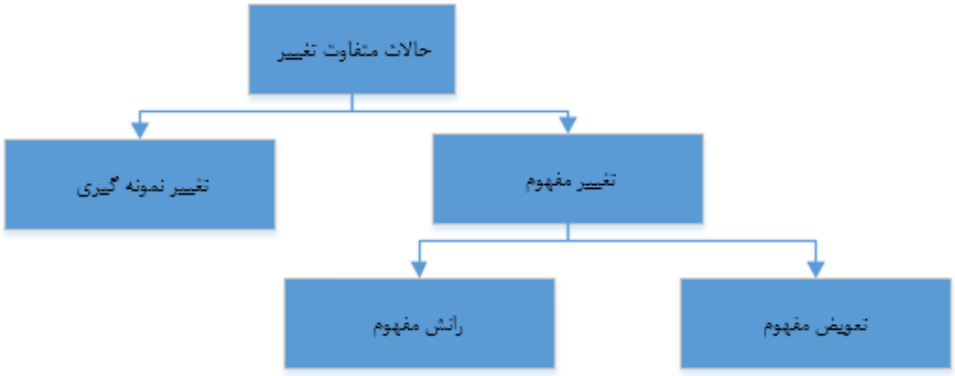
شکل ۱:: انواع مختلف تغییرات در جریان‌های داده‌ای [2]

سه نمودار الف، ب و پ در شکل ۱ سه نوع اصلی رانش‌هایی را که ممکن است در یک متغیر در طول زمان اتفاق بیفتد نمایش می‌دهند. رانش یک‌باره [۱] سریعاً باعث تغییر کلاس متغیر می‌شود. به‌عنوان‌مثال می‌توان به تغییر فصل و تأثیر آن روی فروش اشاره کرد. در رانش مفهوم افزایشی[۲] و تدریجی[۳] مقدار متغیرها به آهستگی در طول زمان تغییر می‌کنند. رانش افزایشی زمانی اتفاق می‌افتد که مقادیر متغیرها به آهستگی در طول زمان دستخوش تغییر می‌شود. در رانش‌های بازگشتی[۴] مفاهیم قبلی پس از مدتی بازمی‌گردند. این تغییرات نسبت به مفهوم پریودیک و یا رفتارهای فصلی و موسمی دارندگان حساب متفاوت می‌باشند چراکه رانش مفهوم غیرقابل‌پیش‌بینی در نظر گرفته می‌شود و زمان قطعی بازگشت آن مشخص نیست. تغییر نوسان گذرا[۵] مربوط به حوادث نادر است که معمولاً برون هشته در نظر گرفته می‌شوند. به‌عنوان‌مثال می‌توان به تراکنش‌های تقلبی در کارت‌های اعتباری اشاره کرد. خش‌ها[۶] نیز تغییرات تصادفی می‌باشند که باید حذف شوند و نباید به‌اشتباه رانش در نظر گرفته شوند.

در مطالعه [11] دسته‌بندی دیگری روی انواع متفاوت رانش مفهوم بیان‌شده است. این دسته‌بندی در شکل ۲ نشان داده‌شده است و حالت کلی‌تری از شکل قبلی می‌باشد.

---

[۱] Sudden drift
[۲] Incremental Drift
[۳] Gradual Drift
[۴] Recurring Drift
[۵] Incremental Drift
[۶] Noise



شکل ۲: انواع متفاوت رانش مفهوم [11]

در این شکل تغییر مفهوم[۱] به تغییر داده‌ها در طول زمان اشاره می‌کند که به دو دسته تقسیم می‌شود؛ رانش مفهوم و تعویض مفهوم[۲]. رانش مفهوم همان تغییر تدریجی و یا افزایشی مفهوم ست و تعویض مفهوم همان رانش یکباره در دسته‌بندی قبل می‌باشد. تغییر نمونه‌گیری[۳] که با نام رانش مفهوم مجازی[۴] نیز شناخته می‌شود تنها به تغییر در توزیع داده‌ها اشاره می‌کند بدین معنی، که با ثابت ماندن مفهوم، تنها توزیع داده‌ها دستخوش تغییر می‌شود. انواع رانش‌ها جامع و کامل نیستند و درجه واقعی رانش مفهوم، ترکیبی از انواع رانش‌هاست.

الگوریتم PSO [۵]

رویکردهای فرا ابتکاری[۶] امروزه کاربرد بسیاری در شاخه‌های مختلف علم بهینه‌سازی پیداکرده‌اند. مبنای این رویکردها عمدتاً بر اساس نظم یا قواعد موجود در ارگانیسم‌های طبیعی یا برگرفته از دیگر شاخه‌های علوم است. رویکردهای فوق بر خلاف روش‌های دقیق بهینه‌سازی، به دنبال نقاط تا حد ممکن نزدیک به بهینه سراسری می‌باشند به‌طوریکه نظر تصمیم‌گیرنده را تا سطح قابل قبولی برآورده سازد. روش بهینه‌سازی ازدحام ذرات از این دسته الگوریتم‌ها است. PSO با یک گروه از جواب‌های تصادفی (ذره‌ها) شروع به کار می‌کند سپس برای یافتن جواب بهینه در فضای مسئله با به‌روز کردن نسلها به جستجو می‌پردازد. هر ذره با دو مقدار X_i و V_i که به ترتیب معرف وضعیت مکانی و سرعت مربوط به i امین ذره هستند تعریف می‌شود. در هر مرحله از حرکت جمعیت، هر ذره با دو مقدار بهترین به‌روز می‌شود [12]. اولین مقدار، بهترین جواب از لحاظ شایستگی[۷] است که تاکنون برای هر ذره به‌طور جداگانه به دست آمده است، این مقدار p_best نامیده می‌شود. مقدار بهترین دیگری که توسط PSO به دست می‌آید. بهترین مقداری است که تاکنون توسط تمام ذره‌ها در میان جمعیت به دست آمده است، این مقدار، بهترین کلی است و g_best نام دارد. بعد از یافتن دو مقدار p_best و g_best، هر ذره سرعت و مکان جدید خود را طبق روابط زیر به‌روز می‌کند:

---

۱ Concept Change
۲ Concept Shift
۳ Sampling changes
۴ Virtual Drift
۵ Particle Swarm Optimization
۶ meta-heuristic
۷ Fitness

$$V_i^{k+1} = W * V_i^k + C_1 * rand * (p\_best_i - S_i) + C_2 * rand * (g\_best_i - S_i) \quad (۱)$$

$$S_i^{k+1} = S_i^k + V_i^{k+1} \quad (۲)$$

$$W = W_{\max} + \frac{(W_{\min} - W_{\max})}{iter\ \ Max} * iter \quad (۳)$$

که لیست متغیرها مطابق جدول ۱ است.

جدول ۱: لیست متغیرهای الگوریتم PSO

سرعت ذره i در تکرار k آم	V_i^k
سرعت ذره i در تکرار k+1 آم	V_i^{k+1}
وزن اینرسی	W
ضریب وزنی	$C_1 = C_2$
مکان فعلی ذره i در تکرار k آم	S_i^k
مکان فعلی ذره i در تکرار k +1 آم	S_i^{k+1}
شماره تکرار فعلی	iter
ماکزیمم تعداد تکرار	iter Max
بهترین جواب ذره i ازلحاظ شایستگی	p_best_i
بهترین جواب گروه ازلحاظ شایستگی	g_best_i
مقدار اولیه برای وزن اینرسی	$W_{\max}$
مقدار نهایی برای وزن اینرسی	$W_{\min}$

شرط توقف را می‌توان به چند صورت تعریف کرد، مثلاً بر روی حداکثر تعداد تکرار، یا اینکه ماکزیمم تغییر در بهترین شایستگی برای دو تکرار متوالی کمتر از تلورانس تعریف‌شده باشد.

$$\left| g\_best^{(k+1)} - g\_best^{(k)} \right| \le \varepsilon \quad \text{Where} = \varepsilon \ 0.001 \quad (۴)$$

رابطه ۱ شامل جمع سه عبارت است که عبارت اول نسبتی از سرعت جاری ذره است و نقش آن شبیه الگوریتم مومنتوم در شبکه‌های عصبی است، به همراه عبارت دوم که متناسب با تفاضل مکان پرنده با بهترین موقعیت قبلی آن و عبارت سوم که تفاضل مکان آن با بهترین جواب در میان کل جمعیت است سبب هدایت سرعت جدید ذره به سمت جواب بهینه می‌شوند.

همگرایی مسئله به پارامترهای PSO مانند w و C_1 و C_2 وابسته است. مقادیر C_1 و C_2 را معمولاً برابر ($C_2=C_1$) و در بازه [0-2] در نظر می‌گیرند. ضریب اینرسی وزنی هم با تغییر توسط رابطه ذکرشده باعث همگرایی خواهد شد؛ که به‌صورت دینامیک در بازه [0.2-0.8] تعریف می‌شود.در این پژوهش روش پنجره‌ای برای به‌روزرسانی پروفایل مشتریان با استفاده از الگوریتم PSO پیاده‌سازی شده است.

مدل پنجره‌ای

رانش مفهوم در طول زمان نتایج طبقه‌بندی را تغییر می‌دهد که این موضوع به دلیل وقوع تغییرات در مفهوم و الگوی داده‌ها است. [13] در چنین سناریوهای یادگیری، داده‌ها به‌طور پیوسته (به‌صورت دسته‌ای) وارد می‌شوند و در طول زمان، تابع اصلی تولیدکننده اطلاعات ممکن است تغییر کند. یادگیری در چنین شرایطی مستلزم آن است که طبقه بند بتواند به چنین تغییراتی پاسخ دهد، درحالی‌که اطمینان حاصل می‌کند که همه دانش و اطلاعات گذشته را حفظ می‌کند.

پروفایل تراکنشی بر اساس ارتباط مشخصه‌های تراکنش‌ها ساخته می‌شود و ازآنجایی‌که توزیع داده‌ها ممکن است در طول جریان داده‌ها تغییر یابد، روش‌های مبتنی بر پنجره کشویی تنها توجه خود را به داده‌هایی معطوف می‌نمایند که اخیراً مشاهده شوند. برخی مجموعه‌های داده‌ای ممکن است در بخشی از جریان داده‌ای بسیار پرتکرار باشد اما ممکن است با ورود داده‌های جدید و افزایش حجم داده‌ها همان مجموعه داده‌ای در بخش دیگری از داده‌ها دارای تعداد بسیار کم و یا حتی صفر باشد. به‌عنوان‌مثال سبد خرید یک فروشگاه را در نظر بگیرید. یک کالای خاص مانند بستنی ممکن است در یک‌فصل خاص و فصل گرما بسیار زیاد بفروش برسد و دارای تعداد تکرار بسیار زیادی درفروش روزانه یک سوپرمارکت باشد اما با رسیدن به فصل سرما به‌تدریج از تعداد تکرار آن کاسته شده و به یک کالای کم تکرار یا حتی با تکرار صفر تبدیل شود. برای رفع این مشکلات الگوریتم‌های مبتنی بر پنجره کشویی معرفی گردیده‌اند که توجه خود را به بخشی از داده‌ها معطوف می‌سازند و داده‌های پرتکرار قبلی که جزء داده‌های اخیر نیستند با کاهش تعداد تکرار مواجهه می‌شوند، به‌طوری‌که رفته‌رفته به اقلام داده‌ای غیر پرتکرار تبدیل می‌شوند و از لیست داده‌هایی که در هر پنجره ارزیابی می‌گردند حذف می‌شوند.

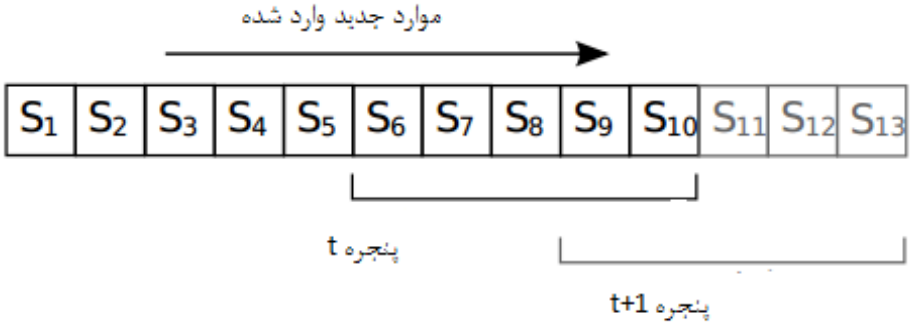
اندازه پنجره مناسب برای یادگیری در این دسته از راهکارها مصالحه‌ای است بین دقت و سرعت؛ درصورتی‌که اندازه پنجره کوچک انتخاب شود یادگیرنده سریعاً به تغییرات پاسخ می‌دهد ولی ممکن است دقتش را به علت زمان کوتاه پایداری و عدم حضور نمونه کافی برای یادگیری از دست دهد. از طرف دیگر اندازه بزرگ پنجره، افزایش دقت را به دنبال خواهد داشت ولی سریعاً به تغییرات پاسخگو نخواهد بود.

روش‌های مبتنی بر پنجره کشویی تنها توجه خود را به داده‌هایی معطوف می‌نمایند که اخیراً مشاهده می‌شوند. چالش‌های اصلی این روش که بر دقت آن تأثیرگذار است شامل موارد زیر است: ۱-آموزش طبقه بندها ۲- ایجاد یک گروه که ترکیبی از داده‌های اخیر با داده‌های تأخیری است و آموزش طبقه بند بر اساس این گروه ۳- انتخاب طول پنجره ۴-استفاده از پنجره‌های کشویی پویا یا ثابت.

یکی از روش‌های کلاسیک برای مقابله با مسئله رانش مفهوم این است که یک پنجره حاوی تعداد M نمونه اخیر که برای آموزش یا به‌روزرسانی طبقه‌بندی استفاده می‌شود، انتخاب می‌شود. با استفاده از این روش، طبقه بند "نمونه‌های آموزش قدیمی" را فراموش خواهد کرد که نشان‌دهنده یک مفهوم قدیمی بوده، درنتیجه حاوی اطلاعات متناقض است. علاوه بر سادگی آن، روش‌های مبتنی بر پنجره، این پرسش را مطرح می‌کند که سایز M چقدر باید باشد که عملکرد خوبی داشته

باشد. یک پنجره کوچک می‌تواند یک سیستم را با واکنش سریع به تغییرات ایجاد کند، اما تعداد کم داده‌های آموزشی ممکن است منجر به کاهش دقت طبقه‌بندی درزمانی که مفهوم پایدار است، شود. یک جایگزین می‌تواند تعریف یک پنجره بزرگ باشد که یک طبقه بند پایدار و آموزش‌دیده ایجاد می‌کند که در زمان‌های تغییر مفهوم به‌آرامی تطبیق یابد [14] و [15]. این سازش بین واکنش سریع و عملکرد خوب در مناطق پایدار با تنظیم پارامترهای فراموش کردن می‌تواند به‌عنوان معضل پایداری-انعطاف پذیری[1] باشد.

شکل ۳، ایده اساسی روش‌های پنجره‌ای را نشان می‌دهد، جایی که هر Si نشان‌دهنده یک نمونه آموزشی (دسته‌ای) است و هرچه مقدار i بیشتر باشد، نمونه جدیدتر است. اندازه پنجره که در شکل استفاده می‌شود، برابر با ۵ است، جایی که فقط نمونه‌های داخل پنجره فعلی برای ساخت یا به‌روزرسانی سیستم طبقه‌بندی استفاده می‌شود. در زمان t، نمونه‌های S6 تا S10 در پنجره فعلی قرار دارند، اما در زمان t + 1 نمونه‌هایی از S11 تا S13 (نمونه‌های خاکستری) می‌رسند، بنابراین نمونه‌های S6 تا S8 از مجموعه آموزشی فعلی حذف‌شده و به سمت پنجره جدید حرکت می‌کنیم.



شکل ۳: مدل پنجره کشویی

خانواده الگوریتم‌های FLORA یکی از اولین روش‌های تحت نظارت پیشنهادشده برای مقابله با رانش مفهوم با استفاده از یک پنجره کشویی[2] است [14]. FLORA اصلی شامل پنجره‌ای با طول ثابت است که هر بار که یک نمونه جدید وارد می‌شود، قدیمی‌ترین آن حذف می‌شود و مدل با استفاده از داده‌های جاری به‌روز می‌شود. در FLORA2 قابلیت انطباق اندازه پنجره اضافه‌شده است به‌طوری‌که اگر الگوریتم یک ناحیه پایدار را تشخیص دهد، طول پنجره را اضافه کرده و اگر یک منطقه در حال تغییر را شناسایی کند، طول پنجره را کاهش می‌دهد. در [16] هولتن و همکاران الگوریتم یادگیرنده درخت تصمیم‌گیری بسیار سریع (CVFDT) سازگار با مفهوم[3] که توسعه‌یافته الگوریتم یادگیرنده درخت بسیار سریع (VFDT) است، ارائه داده‌اند، قابلیت به‌روز نگه‌داشتن درخت را با یک پنجره حاوی M نمونه تحت نظارت را دارد. هر بار که یک نمونه تحت نظارت جدید در دسترس است، آمار گره‌های درخت با توجه به نمونه جدید اضافه‌شده و با توجه به نمونه قدیمی که از پنجره فعلی حذف‌شده است، به‌روز می‌شود.

مسئله اندازه بهینه پنجره در [15] موردبررسی قرارگرفته است به‌طوری‌که معادله‌ای برای تخمین اندازه بهینه پنجره برای رانش مفهوم ناگهانی ارائه‌شده است. در روش پیشنهادشده در [17] یک پنجره کشویی مجموعه داده‌های آموزشی را مشخص

[1] Stability-Plasticity
[2] Sliding Window
[3] Concept-adapting very fast decision tree learner

می‌کند. هر بار که پنجره حرکت می‌کند، یک شبکه عصبی با استفاده از الگوریتم مبتنی بر PSO بازآموزی می‌شود، یعنی الگوریتم PSO مجدداً مقداردهی شده و طور کامل جستجو را پس از حرکت پنجره مجدداً شروع می‌کند.

روش تحقیق

ویژگی‌های انتخابی شامل مبلغ تراکنش، شماره کارت، زمان (ساعت و تاریخ انجام تراکنش) است که در بازه یک سال و نیم از تراکنش‌های کارت مشتریان یکی از بانک‌های خصوصی کشور استفاده‌شده است.

برای در نظر گرفتن بازه مناسب پایداری مفهوم همچنین می‌توان از راهکارهای مبتنی بر پنجره و مفهوم الگوهای پایدار موجود در تراکنش‌های مالی بهره برد. با توجه به اهمیت سرعت تشخیص در حوزه هدف می‌توان طول مناسب برای پایداری مفهوم را با توجه به دارنده کارت مشخص کرد. در عمل محاسبه کارایی با این روش و به ازای مقادیر مختلف پارامتر برای تعداد بالای دارندگان حساب ازلحاظ زمان، بسیار پرهزینه است. به‌علاوه مقادیر متفاوت این پارامتر برای آزمایش نیز غالباً به‌طور تخمینی وارد می‌شوند که مناسب نیست. همچنین رانش مفهوم ممکن است مقدار تخمینی مناسب برای این متغیر را در طول زمان تغییر دهد. در این پژوهش پارامترهایی که تحت تأثیر رانش قرار می‌گیرند به‌طور پویا بر اساس داده‌های موجود تعیین و هشدار تولید می‌شود. در اینجا با استفاده از روش پیشنهادی، طول پنجره بهینه برای مشتریان مختلف محاسبه می‌شود. درواقع هدف اصلی، محاسبه طول پنجره بهینه برحسب تعداد تراکنش به‌منظور به‌روزرسانی پروفایل مشتری به‌طوری‌که به‌دقت حداکثری دست‌یابیم.

در ادامه کارهای انجام‌شده به‌منظور پیاده‌سازی مدل پیشنهادی به‌صورت مختصر شرح داده‌شده است:

۱- ابتدا مقدار اولیه برای طول پنچره در نظر می‌گیریم و بر اساس همین عدد کل تراکنش‌ها به بخش‌های مساوی تقسیم می‌کنیم که طول هر بخش نشان‌دهنده طول پنجره است.
۲- بر اساس طول پنجره مرحله قبل تعداد داده‌های آموزش و آزمایش مشخص می‌شود و این داده‌ها به‌صورت تصادفی انتخاب می‌شوند.
۳- در این مرحله جمعیت تصادفی در بازه داده‌های آموزش توسط الگوریتم PSO تولید می‌شود.
۴- در این مرحله بهینه‌سازی با استفاده از الگوریتم PSO انجام می‌شود که تابع هدف تعریف‌شده مطابق رابطه زیر است.

$$Minimizing \qquad \sum_{i=1}^{N} \left| \frac{particle - learningData(i)}{N} \right| \qquad (۵)$$

که در آن:

Particle: هریک از اعضای جمعیت تولیدشده توسط الگوریتم PSO

learningData: داده‌های آموزشی که در هر پنجره معادل ۸۰٪ کل داده‌ها است.

N: تعداد تراکنش‌ها داده آموزش در هر بخش

۵- در این مرحله، داده‌های آزمایش که ۲۰٪ داده‌های هر بخش است برای محاسبه دقت به الگوریتم داده می‌شود و به ازای هر بخش دقت محاسبه می‌شود و میانگین دقت ذخیره می‌شود.

۶- مقدار دیگری برای طول پنجره در نظر گرفته و کلیه مراحل بالا را تکرار می‌کنیم.

۷- درنهایت بالاترین دقت محاسبه‌شده و طول پنجره متناظر آن به‌عنوان جواب نهایی مسئله خواهد بود.

۸- از مقادیر محاسبه‌شده پروفایل مشتری به‌صورت زیر ساخته می‌شود.

Customer Profile = [Global Best, Personal Best, Window Size]

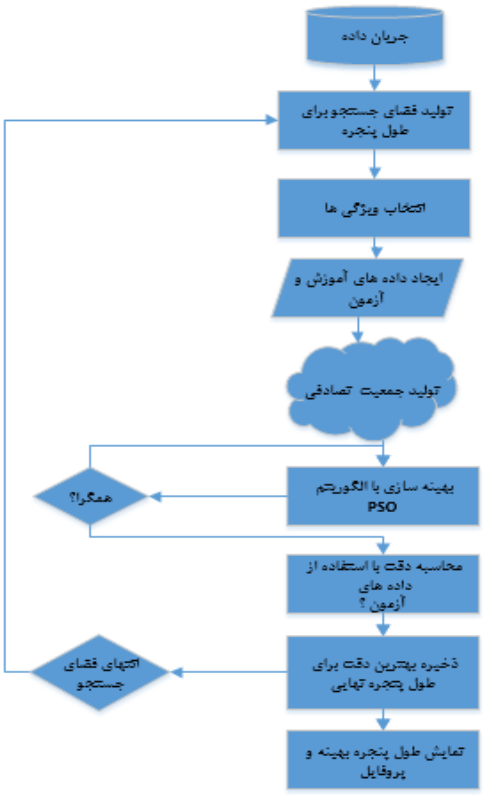
کمترین مقدار تابع هدف به‌عنوان Global Best و عضو جمعیت معادل آن به‌عنوان Personal Best است.

لازم به ذکر است پروفایل ساخته‌شده تنها مشخصات مربوط به مبلغ تراکنش مشتری را شامل می‌شود. به‌منظور لحاظ نمودن پارامتر فاصله زمانی بین تراکنش‌های مشتری به‌عنوان مشخصات زمان، در پروفایل مشتری از دو پارامتر زیر استفاده‌شده است:

۱- حداکثر تعداد تراکنش‌های مشتری در فاصله زمانی یک ساعت

۲- حداکثر تعداد تراکنش‌های مشتری در فاصله زمانی ۲۴ ساعت

درنهایت پروفایل مشتری بر اساس مورد مرحله ۸ و دو پارامتر ذکرشده، ساخته می‌شود. به‌منظور شناسایی تراکنش‌های نابهنجار مشتری، تراکنش‌های جدید با هر یک از پارامترهای مربوط به پروفایل مقایسه می‌شود و در صورت نقض هر یک از این موارد هشدار تولید می‌شود.



شکل ۴: مدل روش پیشنهادی

یافته‌ها و نتایج

در هنگام ارزیابی عملکرد روش‌های طبقه‌بندی در محیط‌های ایستا، برخی از روش‌های معمول ممکن است شامل معیارهای کلاسیک مانند دقت، پارامترهای ماتریس اغتشاش و تجزیه‌وتحلیل منحنی ROC باشد. علیرغم اهمیت آن‌ها، این معیارهای کلاسیک به‌طور کامل نمی‌تواند کیفیت روش‌هایی را که با مشکلات رانش مفهوم مواجه هستند، به‌طور کامل نشان دهد، زیرا عملکرد این روش‌ها در طول زمان متفاوت است. [18] بنابراین معیار دقت متوسط مطابق رابطه زیر نیز در نظر گرفته می‌شود و نمودار تغییرات متوسط دقت در نتایج ارائه می‌گردد.

Precision= TP/ (TP+FP)

در برچسب‌گذاری تراکنش توسط روش پیشنهادی یکی از چهار حالت زیر پیش می‌آید:

TP: تراکنش نابهنجار که به‌درستی نابهنجار شناسایی‌شده است.
FP : تراکنش بهنجاری که به‌اشتباه نابهنجار شناسایی‌شده است.
TN : تراکنش بهنجاری که به‌درستی بهنجار شناسایی‌شده است.
FN : تراکنش نابهنجار که به‌اشتباه بهنجار شناسایی‌شده است.

در حالت بهینه روش‌های کشف تقلب باید بتوانند در حضور رانش مفهوم در عین اینکه تعداد FP را در حداقل خود نگه‌دارند TPرا کاهش ندهند. درصورتی‌که FP در سیستم‌های کشف تقلب در حد قابل قبولی نگاه داشته نشود، نارضایتی مشتری را به دنبال خواهد داشت. به‌علاوه اخطارهای حاصل از سامانه کشف تقلب برای تحلیل‌های بعدی برای بخش تقلب سازمان مربوطه فرستاده می‌شود و موارد مشکوک برای تأیید و یا عدم‌تأیید تقلب بودن بررسی می‌شوند. درنتیجه تعداد اخطارها باید در سطحی نگاه داشته شوند که قابل اداره کردن باشند.

به‌منظور تشخیص انواع رانش مفهوم از دودسته مشتریان با تعداد تراکنش بالا و تراکنش پایین استفاده‌شده است و برخی نتایج آن در ادامه نشان داده‌شده است.

نتایج مطالعه الگوهای رفتاری عمومی مشتریان و متقلبان شامل برخی موارد است که در ادامه مطرح می‌شود.

- مشتریان غالباً از الگوهای رفتاری پایداری تبعیت می‌کنند. الگوهای رفتاری کاملاً تصادفی به‌ندرت در میان مشتریان دیده می‌شود.
- تغییرات رفتاری مشهود موسمی در رفتارهای مشتریان دیده می‌شود. این تغییرات مربوط به آب‌وهوا، تعطیلات و یا رویدادهای مهم است مثل تغییرات رفتاری در نوروز.
- الگوهای دوره‌ای هفتگی و ماهیانه قابل‌توجهی در رفتارهای مشتریان دیده می‌شود.
- مشتریان را می‌توان به سه دسته کم‌مصرف، با مصرف متوسط و پرمصرف تقسیم کرد. مشتریان پرمصرف مشتریانی هستند که درصد بالایی از خریدهایشان نرخ بالایی دارد یا تعداد تراکنش بالایی دارند.

در این تحقیق ما الگوهای متفاوت رانش مفهوم را اعمال کرده‌ایم بدین معنی که مشتریان به‌مرورزمان و با تغییر شرایط محیطی رفتارشان تغییر می‌کند. در رانش مفهوم، رفتار مشتری به این صورت در نظر گرفته می‌شود که مقادیر تفاوت معنادار در رفتار مشتری با توجه به مبلغ تراکنش محاسبه می‌شود. همچنین حداکثر تعداد تراکنش‌های مشتری در فاصله زمانی یک ساعت و حداکثر تعداد تراکنش‌های مشتری در فاصله زمانی ۲۴ ساعت در مقایسه با رفتار بهنجار اولیه تغییر می‌کنند. این تغییر با توجه به اینکه کدام‌یک از انواع رانش مفهوم در رفتار دارنده حساب رخ‌داده است، اعمال می‌شود.

رانش مفهوم به سه صورت تغییر می‌کند.

- رانش یک‌باره؛ مقادیر ذکرشده با فاکتوری تغییر می‌کنند که توزیعی بسیار متفاوت با توزیع اولیه برای مشتری تولید شود.
- رانش افزایشی؛ مقادیر ذکرشده با فاکتوری کوچک تغییر می‌کنند تا توزیع داده‌های تولیدی نزدیک به توزیع اولیه برای مشتری باشد.
- رانش تدریجی؛ برای جایگزین شدن تدریجی رفتار جدید توزیع برخی از داده‌ها ثابت می‌ماند و سایر داده‌ها با فاکتوری که توزیع متفاوتی با توزیع اولیه ایجاد کند تغییر می‌کنند. این روند ادامه می‌یابد تا جای که توزیع تمامی داده‌ها متفاوت از توزیع اولیه می‌شود.

داده‌های برای ارزیابی این روش شامل تراکنش‌های هریک از دارندگان کارت است. تراکنش‌های یک سال مشتری (از نیمه سال ۹۶ تا نیمه سال ۹۷) برای هریک از مجموعه‌های آموزش و آزمون ساخته‌شده است. سعی بر این بوده است که رفتارهای متنوع دارندگان کارت پوشش داده شود. در جدول ۱ نمونه ای از داده های بررسی شده در این تحقیق مشاهده می‌شود.

جدول ۱: مشتریان و نوع تراکنش

	مشتریان پرمصرف	تعداد تراکنش های نابهنجار	مشتریان کم‌مصرف	تعداد تراکتش های نابهنجار
رانش مفهوم یک‌باره	۱۲۰۲-کارت ۱	۸۲	۵۱۹-کارت ۴	۴۱
رانش مفهوم تدریجی	۱۱۲۳-کارت ۲	۶۸	۶۴۷- کارت ۵	۵۲
رانش مفهوم افزایشی	۱۰۹۵-کارت ۳	۷۱	۵۷۱- کارت ۶	۴۷

جدول ۲: نتایج مدلسازی برای رانش یک‌باره در مشتری پرمصرف

		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۶۸	۱۹
	Negative	۱۴	۱۱۰۱

جدول ۳: نتایج مدلسازی برای رانش تدریجی در مشتری پرمصرف

		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۵۳	۳۲
	Negative	۱۵	۱۰۲۴

جدول ۴: نتایج مدلسازی برای رانش افزایشی در مشتری پرمصرف

		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۵۹	۲۱
	Negative	۱۲	۱۰۰۳

جدول ۵: نتایج مدل در مشتریان پرمصرف

	F measure	دقت(Precision)	صحت(Recall)
رانش مفهوم یک‌باره	۰,۸۰۵۰	۰,۷۸۱۶	۰,۸۳۱
رانش مفهوم تدریجی	۰,۶۹۲۸۱	۰,۶۲۳۵۳	۰,۷۷۹۴۱
رانش مفهوم افزایشی	۰,۷۸۱۴۶	۰,۷۳۷۵	۰,۸۳۰۹۹

جدول ۶: نتایج مدلسازی برای رانش یک‌باره در مشتری کم‌مصرف

		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۳۳	۱۹
	Negative	۸	۴۵۹

جدول ۷: نتایج مدلسازی برای رانش تدریجی در مشتری کم‌مصرف

		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۴۱	۲۷
	Negative	۱۱	۵۶۸

جدول ۸: نتایج مدلسازی برای رانش افزایشی در مشتری کم‌مصرف

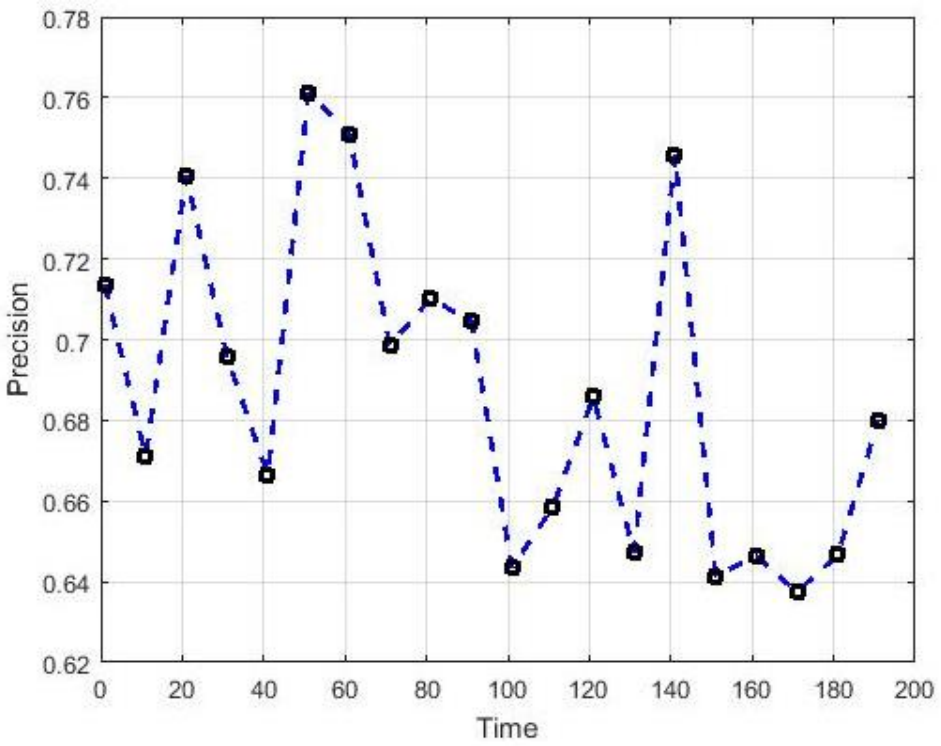
		وضعیت واقعی	
		Positive	Negative
نتیجه طبقه بند	Positive	۳۸	۱۵
	Negative	۹	۵۰۹

جدول ۹: نتایج مدل در مشتریان کم‌مصرف

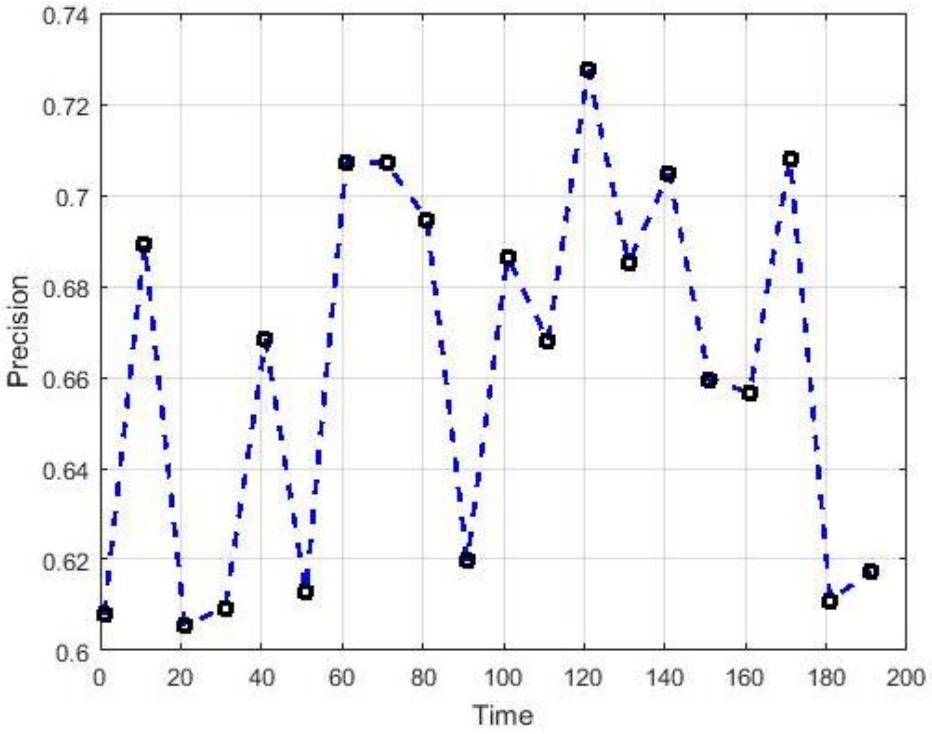
	F measure	دقت(Precision)	صحت(Recall)
رانش مفهوم یک‌باره	۰,۷۸۱۴۶	۰,۸۳۰۹۹	۰,۷۳۷۵
رانش مفهوم تدریجی	۰,۶۸۳۳۳	۰,۶۰۲۹۴	۰,۷۸۸۴۶

رانش مفهوم افزایشی	۰,۷۶	۰,۷۱۶۹۸	۰,۸۰۸۵۱

متوسط تغییرات دقت برحسب زمان برای مشتریان کم‌مصرف و پرمصرف به ترتیب در شکل ۴ و ۵ نشان داده‌شده است.

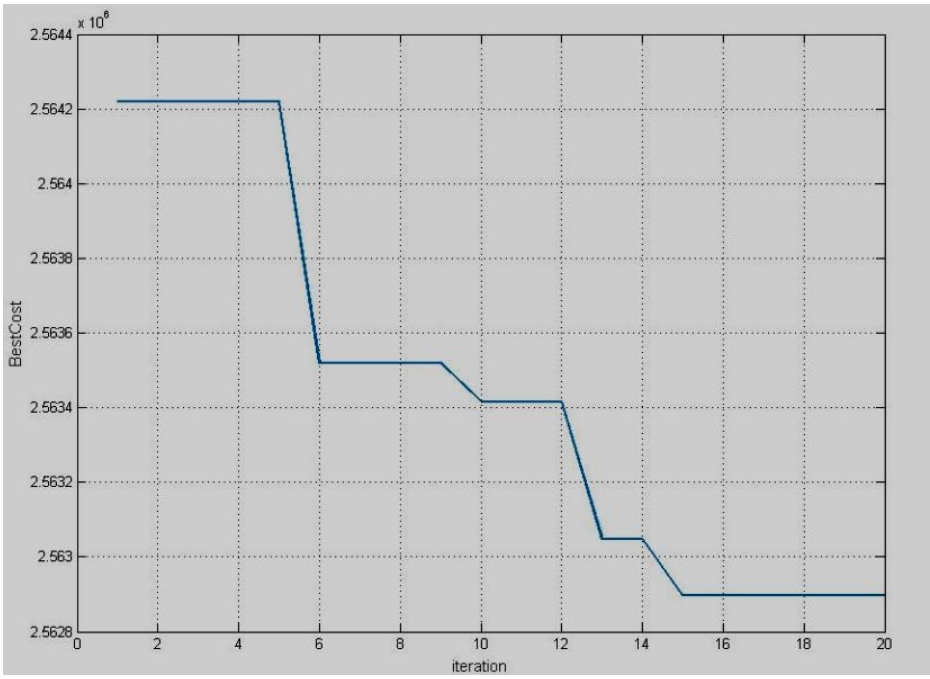


شکل ۵: متوسط تغییرات دقت در مشتریان پرمصرف



شکل ۶: متوسط تغییرات دقت در مشتریان کم‌مصرف

همانطورکه در شکل نشان داده‌شده است برای هر دو مدل مشتریان، متوسط دقت بالای ۶۰ درصد است؛ بنابراین الگوریتم حداقل ۶۰ درصد تراکنش‌های نابهنجار را به‌درستی تشخیص داده است. برای تراکنش‌های با رانش مفهوم یک‌باره دقت تشخیص تراکنش‌های نابهنجار حدود ۸۰ درصد است. بنابراین مدل پنجره کشویی در تشخیص رانش مفهوم یک‌باره نسبت به انواع دیگر از دقت بالاتری برخوردار است. همچنین مطابق نتایج به‌دست‌آمده دقت این روش در تشخیص تراکنش‌های نابهنجار برای مشتریان کم‌مصرف بهتر است.



شکل ۷: منحنی تغییرات BestCost

در شکل، منحنی تغییرات g_best نسبت به تعداد تکرار برای یکی از مشتریان پرمصرف نشان داده‌شده است. مقدار عددی BestCost مطابق رابطه (۵) به‌عنوان تابع fitness در الگوریتم PSO محاسبه می‌شود. کاهش مقدار این عدد همان‌طور که در شکل آمده است، درواقع نشان‌دهنده رسیدن به بهترین جواب برای تابع هدف است. همان‌طور که مشاهده می‌شود مقدار BestCost پس از ۱۶ تکرار مطابق رابطه (۴) همگرا شده است که نشان‌دهنده پایداری الگوریتم است.

جمع‌بندی

در این تحقیق به‌روزرسانی پروفایل مشتریان در سیستم‌های کشف تقلب با در نظر گرفتن رانش مفهوم را هدف قراردادیم و مدلی برای مدیریت رانش مفهوم در رفتار دارندگان کارت در این دسته از سامانه‌ها ارائه دادیم. مدل ارائه‌شده شامل مدیریت رانش مفهوم و به‌روزرسانی پروفایل است. در این تحقیق از مدل پنجره کشویی استفاده شد و نتایج ارزیابی آن بررسی و مقایسه شد. نتایج به‌دست‌آمده نشان می‌دهد که روش پیشنهادی برای انواع مشتریان کم‌مصرف و پرمصرف با رانش‌های مفهوم افزایشی، تدریجی و یک‌باره در تشخیص تراکنش‌های نابهنجار دقت بیش از ۶۰ درصد را دارا است. از نقاط ضعف این چارچوب افزایش مدت‌زمان آموزش و مصالحه بین سرعت و دقت است.

می‌توان در مؤلفه مدیریت رانش مفهوم روش‌ها و مدل‌های متعددی را آزمود. برای مثال می‌توان از روش گروه‌بندی طبقه بندها برای به‌روزرسانی پروفایل مشتریان در حضور رانش مفهوم استفاده نمود. بعلاوه می‌توان از داده‌های شبکه‌های اجتماعی و وب، تعیین تمایل و انگیزه مشتری به تقلب، تحلیل‌های معنایی[1] در ساخت پروفایل مناسب بهره برد.

منابع

[1] Johannes Jurgovsky, Michael Granitzer, Konstantin Ziegler,” Sequence Classification for Credit-Card Fraud Detection”, Expert Systems With Applications,pages 5,2018.

[2] D. Brzezinski, "Mining Data Streams with Concept Drift," Master of Science, Poznan University of Technology, Poland, 2010.

[3] Y. Gmbh and K. G. Co, “Global online payment methods: Full year 2016,” Tech. Rep, 3 2016.

[4] A. Kundu, S. Sural, and A. Majumdar, "Two-stage credit card fraud detection using sequence alignment", in Information Systems Security, Lecture Notes in Computer Science, 2006, Volume 4332/2006, pp. 260-275.

[5] B. Wiese and C. Omlin, "Credit Card Transactions, Fraud Detection, and Machine Learning: Modelling Time with LSTM Recurrent Neural Networks," Innovations in Neural Information Paradigms and Applications, Berlin, Heidelberg: Springer Berlin Heidelberg, vol.247, 2009.

[6] J. Quah and M. Sriganesh, "Real-time credit card fraud detection using computational intelligence," Expert Systems with Applications, vol. 35, no. 4, 2008, pp. 1721-1732.

[7] M.M. Masud, J. Gao, L. Khan, and B. Thuraisingham, "Classification and novel class detection in concept-drifting data streams under time constraints," IEEE Transactions on Knowledge and Data Engineering, vol. 23, no. 6, 2011, pp. 1-37.

[8] Anuj Sharma & Prabin Kumar Panigrahi,A Review of Financial Accounting Fraud Detection based on Data Mining Techniques,International Journal of Computer Applications 39(1), 2007.

[9] W. Gao, "Constructing user behavioral profiles using data-mining-based approach," doctoral dissertation, Dept. Committee on Business Administration, The University of Arizona, Tucson, Arizona, 2005.

[10] D. K. Tasoulis, N. M. Adams, and D. J. Hand. Unsupervised clustering in streaming data. In ICDM Workshops, pages638-642, 2006.

[11] Y. Yang, X. Wu, and X. Zhu, "Mining in Anticipation for Concept Change: Proactive-Reactive Prediction in Data Streams," Data Min. Knowl. Discov., vol. 13, 2006, pp. 261-289.

[12] M. P. Wachowiak, R. Smolíková, Y. Zheng, J. M. Zurada, and A. S. Elmaghraby, “An Approach to Multimodal Biomedical Image Registration Utilizing Particle Swarm Optimization”, IEEE Trans. on Evolutionary Computation, vol 8, No, JUNE 2004.

semantic[1]

[13] L. L. Minku, A. P. White, and Y. Xin, "The Impact of Diversity on Online Ensemble Learning in the Presence of Concept Drift," Knowledge and Data Engineering, IEEE Transactions on, vol. 22, pp. 730-742, 2010.

[14] GAMA, J.; SEBASTIÃO, R.; RODRIGUES, P. P. On evaluating stream learning algorithms. Machine Learning, Springer US, v. 90, n. 3, p. 317–346, 2013. ISSN 0885-6125.

[15] KUNCHEVA, L. I.; ŽLIOBAITE, I. On the window size for classification in changing environments. Intelligent Data Analysis, IOS Press, Amsterdam, The Netherlands, The Netherlands, v. 13, n. 6, p. 861–872, dez. 2009.

[16] HULTEN, G.; SPENCER, L.; DOMINGOS, P. Mining time-changing data streams. In: Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2001. (KDD '01), p. 97–106. ISBN 1-58113-391-X.

[17] RAKITIANSKAIA, A.; ENGELBRECHT, A. Training feedforward neural networks with dynamic particle swarm optimisation. Swarm Intelligence, Springer US, v. 6, n. 3, p. 233–270, 2012. ISSN 1935-3812.

[18] D. Hand, C. Whitrow, N. Adams, P Juszczak, and D. Weston. Performance criteria for plastic card fraud detection tools. Journl of the Operaional Research Society, 59(7):956-962, 2008.