سال دوازدهم، شماره‌های ۴۳ و ۴۴، بهار و تابستان ۱۳۹۹
صص: ۱۰۵_۱۱۷

دو فصلنامه علمی
فناوری اطلاعات و ارتباطات ایران

ارائه یک روش سریع و دقیق برای شناسایی رانش مفهوم با تحلیل سابقه‌ی رویدادها

مهدی یعقوبی* علی سبطی* سهیلا کرباسی*
*استادیار گروه مهندسی کامپیوتر، دانشکده فنی مهندسی گرگان، دانشگاه گلستان، گرگان
تاریخ دریافت: ۱۳۹۹/۰۲/۰۷ تاریخ پذیرش: ۱۳۹۹/۰۸/۲۴
نوع مقاله: پژوهشی

چکیده

در سازمان‌ها و شرکت‌های بزرگ که از سیستم‌های مدیریت فرآیندهای کسب و کار (BPMS) بهره می‌برند، در هر لحظه با توجه به قوانین بالادستی و شرایط بازار، ممکن است در فرآیندهای کسب و کار تغییرات رخ دهد. این تغییرات گاهی به صورت آنی و گاهی به صورت تدریجی روی سیستم اعمال می‌گردد. شناسایی به موقع این تغییرات می‌تواند در تصمیم‌گیری بهتر مدیران سازمان اثر گذار باشد. تجزیه و تحلیل سابقه‌ی رویدادها در این سیستم‌ها، امکان شناسایی تغییرات ایجاد شده در فرآیندهای کسب و کار را به صورت خودکار فراهم می‌کند. به این تغییرات در فرآیندها به اصطلاح رانش مفهوم در فرآیند کسب و کار گفته می‌شود. استخراج رانش مفهوم اشاره دارد به شناسایی محل و نوع تغییراتی که در طول زمان در فرآیندهای کسب و کار یا به طور کلی در سابقه‌ی روبداد رخ داده است. در این مقاله یک روش ابتکاری با معرفی یک تابع فاصله اصلاح شده، برای شناسایی محل و زمان ایجاد رانش مفهوم ارائه می‌شود. آزمایش‌های انجام شده بر روی ۷۲ پایگاه دادِگان موجود در پیشینه‌ی پژوهش که شامل ۶۴۸ رانش مفهوم در ۱۲ نوع مختلف است، نشان می‌دهد روش پیشنهادی ۹۸/۱۸ درصد از رانش‌ها را تشخیص می‌دهد درحالی که روش پیشنهادی نسبت به بهترین روش موجود بسیار سریع‌تر است.

واژگان کلیدی: مدیریت فرآیندهای کسب و کار، شناسایی تغییرات در فرآیند، رانش مفهوم، فرآیندکاوی

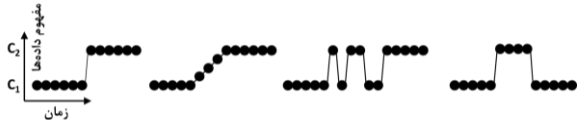
۱- مقدمه

تجزیه و تحلیل سابقه‌ی رویداد[۱] در سیستم های مدیریت فرآیندهای کسب و کار(BPMS) یکی از محبوب‌ترین تحقیقاتی است که در این حیطه انجام می‌شود. یکی از این تحقیقات فرآیند کاوی است. فرآیند کاوی به دسته‌ای از تحقیقات بر روی سابقه‌ی رویداد تمرکز دارد که هدف اصلی آنها استخراج[۲] مدل فرآیند، سازگاری یا تطبیق[۳] مدل استخراج شده با سابقه‌ی رویداد و بهبود و توسعه مدل[۴] موجود می‌باشد[۱] الگوریتم‌های زیادی برای استخراج فرآیند از سابقه‌ی رویداد وجود دارد که از آن دسته می‌توان الگوریتم‌های آلفا، آلفا+، آلفا++، آلفا# ، ابتکاری، نام برد. تکنیک‌های فرآیند کاوی معاصر، عموماً فرآیندها را پایدار در نظر می‌گیرند، در صورتیکه امروزه در واقعیت فرآیندهای کسب و کار در طول زمان و در فواصل مختلف ممکن است دست‌خوش تغییر و تحول شوند. این تغییرات به دلایل مختلفی همچون تغییر در قوانین سازمان‌ها، شرایط فصلی، تغییر در عرضه و تقاضا، تعادل بارِکاری، تعدیل نیرو و یا بروز بلایا و فجایع طبیعی صورت می‌گیرد. این تغییرات ممکن است در طول زمان تاثیرات عمیقی بر روی کارایی فرآیندهای سازمانی بگذارند، به همین دلیل شناسایی محل وقوع و دلیل این تغییرات، برای مدیران سازمان‌ها از اهمیت ویژه‌ای برخوردار است. به این نوع تغییرات در اجرای فرآیندهای کسب و کار که در طول زمان باعث

نویسنده مسئول: مهدی یعقوبی m.yaghoubi@gu.ac.ir

[۱] Event log
[۲] Discovery
[۳] Conformance checking
[۴] Enhancement of process models

تغییراتی در سابقه‌ی رویداد می‌شود، رانش مفهوم[۵]می‌گویند. رانش مفهوم در فرآیند می‌تواند به صورت تدریجی، ناگهانی، افزایشی و مقطعی صورت بگیرد[۲]. در تغییرات ناگهانی، فرآیند جدید جایگزین فرآیند موجود می‌شود و معمولاً در مواقع اضطراری یا با تغییر قوانین صورت می‌گیرد. در تغییرات تدریجی نیز فرآیند موجود توسط فرآیند جدید جایگزین می‌شود اما برخلاف تغییر ناگهانی، در اینجا هر دو فرآیند برای مدت زمانی با هم در حال اجرا هستند و سپس فرآیند پیشین به تدریج کنار گذاشته می‌شود. در تغییرات افزایشی تغییرات به صورت مرحله به مرحله انجام می‌شود و در تغییرات مقطعی، تغییرات برای مقطع زمانی کوتاهی اعمال می‌شود و دوباره به روال قبلی برمی‌گردد. در شکل (۱) انواع رانش‌های مفهوم (تغییرات) بین دو مفهوم (گونه‌ی فرآیند کسب و کار) C_1 و C_2 به تصویر کشیده شده است. شناسایی انواع رانش و تفکیک هر یک از آنها از چالش‌های اصلی در شناسایی رانش مفهوم است. در روش‌های موجود توجه روی شناسایی رانش‌های ناگهانی است و دقت شناسایی در انواع دیگر نسبت به رانش‌های ناگهانی کمتر است.



(الف) مقطعی (ب) تدریجی (ج) افزایشی (د) ناگهانی

شکل ۱- انواع رانش مفهوم از نظر اثر تغییر در سابقه‌ی رویداد در طول زمان [۲]

۱-۱- رانش مفهوم

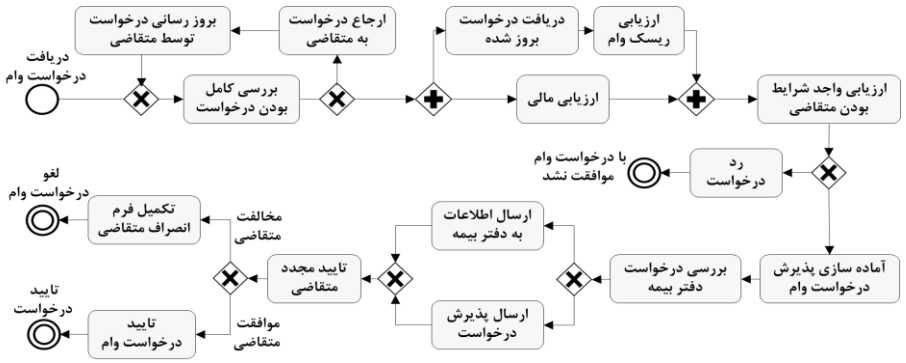
فرآیندکاوی یکی از مسائل مورد توجه در BPMS است که هدف آن استخراج مدل فرآیند از سابقه‌ی رویداد است. مدل فرآیند شامل ساختارهای هم‌زمانی، انتخابی، توالی و حلقه‌ها است که اجرای آن به شکل دنباله‌ای از رویدادها در سابقه‌ی رویداد ذخیره می‌شود. در حالی که شناسایی و استخراج رانش مفهوم مسئله‌ی پیچیده‌تری است و هدف آن کشف تغییرات رخ داده در ساختار مدل فرآیند است که باید با آنالیز و تحلیل داده های ذخیره شده در سابقه‌ی رویداد کشف شود. استخراج رانش مفهوم نیز یکی از مسایل پر طرفدار در شاخه‌های مختلف هوش مصنوعی است و در داده کاوی به عنوان یک مسئله رایج هم به صورت یادگیری با نظارت و هم بدون نظارت مورد مطالعه قرار گرفته است [۶-۳] با این حال، این مسایل به عنوان بخشی از مراحل فرآیند کاوی مورد توجه قرار نگرفته است. اگر چه تجربیات بدست آمده در داده کاوی و یادگیری ماشین در فرآیند کاوی قابل استفاده است، اما پیچیدگی های ساختار فرآیند مانند ساختارهای انتخابی[۶] حلقه، هم روندی، انصراف[۷] و انتخاب‌های مشروط[۸](وابسته به انتخاب‌های قبلی در اجرای فرآیند) چالش‌های زیادی را در استخراج رانش مفهوم از سابقه‌ی رویداد ایجاد می‌کند.

همانطور که گفته شد، فرآیندها ممکن است به منظور تطبیق با تغییرات محیطی و شرایط جدید، دچار تغییر شوند. برای مثال، شرکت‌های مسافرتی در فصول مختلف در فرآیند فروش خود دچار تغییر می‌شوند. بنابراین سازمان‌ها برای رسیدن به عملکرد بهتر باید بتوانند به خوبی با تغییرات محیطی تطبیق یابند. رانش مفهوم زمانی رخ می‌دهد که فرآیند در هنگام اجرا دچار تغییر شود. نیاز به مدیریت این تغییرات، محققان را بر آن داشته که روش‌ها و ابزارهای گوناگونی جهت تحلیل کسب و کار و شناسایی تغییرات فرآیند ارائه دهند. تا کنون روش‌های مختلفی برای شناسایی رانش مفهوم در فرآیندهای کسب و کار با استفاده از سابقه‌ی رویداد ارائه شده است که هر کدام از آن‌ها قادر به شناسایی دسته خاصی از تغییرات می‌باشند. تحلیل تغییرات در فرآیندهای سازمانی از اهمیت بسیار بالایی برخوردار است، زیرا می‌تواند به شناسایی تغییرات و بهبود اجرای فرآیندها، یک دیدگاه صحیح از اجرای فرآیندها در هر بازه زمانی بدست آورد و در نهایت باعث بهبود عملکرد سیستم خواهد شد.

۱_۲ عوامل ایجاد رانش مفهوم در فرآیندها

ماراجی و همکاران [۷] عوامل تغییر فرآیند را معرفی کردند و این عوامل را با ۱۲ الگو متفاوت طبقه بندی کردند، در جدول(۱) لیست این عوامل ذکر شده است. ماراجی این عوامل را در سه گروه درجی(I)، انتخابی(O) و تغییر ترتیب(R) دسته‌بندی کرد و بر هر یک از عوامل جدول (۱) چهار فایل پایگاه دادِگان با اندازه های ۱۰۰۰۰، ۷۵۰۰، ۵۰۰۰ و ۲۵۰۰ پیمایش[۹]با استفاده از تکنیک های شبیه سازی روی فرآیند *درخواست وام* (شکل(۲)) ایجاد کرد و ۶ فایل دادِگان ترکیبی نیز از ترکیب عوامل هر دسته ایجاد کردند که در مجموع ۷۲ فایل دادگان (۱۲ عامل اصلی و ۶ عامل ترکیبی در ۴ اندازه مختلف) توسط

[۵] Concept drift
[۶] Choice
[۷] Cancelation
[۸] Non-free Choice
[۹] Trace



شکل ۲- فرآیند درخواست وام جهت ایجاد پایگاه دادگان[۷]

ماراجی و همکاران ایجاد شد. در هر فایل دادگان، ۹ رانش مفهوم به صورت عمدی ایجاد شده است که در مجموع ۶۴۸ رانش مفهوم را شامل می‌شود. پایگاه دادگان ماراجی در مقالات بعدی مورد توجه نویسندگان و محققان قرار گرفت و محققان برای آزمایش روش‌های خود از آن استفاده کردند.

ردیف	شرح	اختصار	گروه
۱	حذف یا ایجاد یک بخش در فرآیند	re	درجی (I)
۲	قرار دادن یا خارج کردن یک بخش در انشعاب شرطی	cm	
۳	تکرار دوتایی یک بخش	cp	
۴	قرار دادن یا خارج کردن یک بخش در انشعاب موازی	pm	
۵	تعویض یک بخش از فرآیند	rp	
۶	جابجایی دو بخش	sw	
۷	ایجاد حلقه رو یک بخش یا حذف حلقه	lp	انتخابی (O)
۸	ایجاد پرش از روی یک بخش یا حذف پرش	Cb	
۹	تغییر در فراوانی انشعاب انحصاری	Fr	
۱۰	همگام سازی دو بخش	cd	تغییر ترتیب(R)
۱۱	تغییر دو بخش از انتخاب انحصاری به توالی و بالعکس	cf	
۱۲	تغییر دو بخش با اجرای موازی به توالی و بالعکس	pl	

جدول ۱- عوامل ایجاد رانش مفهوم در فرآیندها[۷]

در این مقاله، از یک روش رایج جهت استخراج رانش مفهوم از سابقه‌ی رویداد استفاده می‌شود که با حرکت دو پنجره با اندازه یکسان و ثابت اطلاعات دو بردار ویژگی را از دو ناحیه کنار هم در سابقه‌ی رویداد استخراج می‌کند، سپس به یک روش ابتکاری دو بردار ویژگی برای هر پنجره ایجاد می‌شود و از یک تابع فاصله‌ی جدید(معرفی شده در این مقاله) با عنوان symgTest[۱۰] استفاده می‌شود که منجر به استخراج بهتر رانش مفهوم در فرآیندهای کسب و کار می‌شود.

۲- پیشینه پژوهش

سیستم‌های مدیریت فرآیندهای کسب و کار به گونه‌ای طراحی شده‌اند که امکان تغییر در فرآیندهای کسب و کار را به سادگی برای کاربران فراهم می‌کند. این تغییرات به صورت مداوم در پاسخ به عوامل خارجی مانند اسناد بالادستی، قوانین و مقررات جدید، شرایط بازار، تغییر در تقاضای مشتریان، تغییرات فصلی و سال مالی به وجود می‌آید. برخی از این تغییرات با اطلاع قبلی و مستند می‌باشد، ولی برخی از آنها بدون آگاهی و گاهی به‌صورت سهوی به موجود می‌آیند و بعد از گذشت مدتی به عنوان یک تغییر هنجار در سیستم باقی می‌مانند. برخی از این تغییرات ممکن است در کد منبع سیستم‌های اطلاعاتی خارج از مدل فرآیند ایجاد شود که ناشی از نبودن امکان یا عدم پشتیبانی سیستم مدیریت فرآیند باشد. اگرچه پژوهش‌های زیادی برای شناسایی فرآیندها از سابقه‌ی رویداد انجام شده است[۸، ۹] ولی اکثر پژوهش‌های انجام شده فرض می‌کنند فرآیند در طول دوره‌ای که اجرا شده است تغییری در آن رخ نداده است [۱۰] در حالی که عبارت رانش مفهوم در فرآیند یک اصطلاح شناخته شده است و شناسایی آن قبل از استخراج فرآیند می‌تواند باعث بهبود فرآیند کاوی و ارائه فرآیندهای مستخرج شده قابل انطباق‌تر[۱] شود [۱۴-۱۱].

گوندر و همکاران [۱۵] برای شناسایی تغییرات در فرآیند از سابقه‌ی تغییرات در مدل فرآیند استفاده شده است این بدان معناست که نویسندگان از یک دانش بالا دستی برای شناسایی تغییرات استفاده کرده‌اند و در واقع امکان شناسایی تغییراتی که

[۱] Symmetric Goodness of fit Test

[۱] Fitness

مستند سازی نشده است در روش آنها وجود ندارد و صرفاً تغییرات مدل فرآیند (نه اثر تغییرات در اجرای فرآیند) مورد بررسی قرار گرفته است.

بوز (Bose) و همکاران [۱۶] تنها با بررسی سابقه‌ی رویدادها توانستند محل رانش مفهوم در فرآیند را شناسایی کنند. آنها برای شناسایی محل رانش هم از ویژگی‌های محلی و هم از ویژگی‌های عمومی رویدادها استفاده کردند، با این حال قادر به شناسایی همه انواع رانش مفهوم در فرآیند نبودند. از طرف دیگر در روش آنها لازم بود، کاربر پارامتری را تحت عنوان اندازه‌ی پنجره تنظیم کند تا رانش مفهوم به درستی شناسایی شود در حالی که تنظیم این پارامتر نیاز به یک دانش اولیه در مورد سابقه‌ی رویداد ورودی دارد، در روش آنها میزان دقت نتایج بسیار به پارامتر اندازه‌ی پنجره وابسته بود.

ماراجی و همکاران [۷] برای رفع مشکل بوز از یک پنجره‌ی انطباقی برای شناسایی رانش مفهوم استفاده کردند. در روش آنها از آزمون آماری G-test بر روی دو پنجره استفاده می‌شود، یک پنجره، به عنوان پنجره‌ی ارجاع و دیگری به عنوان پنجره‌ی تشخیص، در هر کدام، تعدادی از پیمایش‌ها[۲]موجود در سابقه-ی رویداد مورد بررسی قرار می‌دهند و تفاوت در توزیع آماری داده‌ها ملاک اصلی تشخیص رانش مفهوم در پنجره‌ی تشخیص است.

مارتجوشو (Martjushev) و همکاران [۱۴] مانند ماراجی از ایده پنجره انطباقی استفاده کردند و الگوریتمی را تحت عنوان Change Point معرفی کردند. آنها الگوریتم خود را برای شناسایی رانشهای مفهوم تدریجی توسعه دادند. سیلیگر (Seeliger) و همکاران [۱۷] از الگوریتم Change Point استفاده کردند و برای استخراج ویژگی‌های پنجره های ارجاع و تشخیص از ویژگی‌های گراف مدل فرآیند استخراج شده در هر پنجره استفاده کرد و برای آزمایش الگوریتم خود از پایگاه دادگان ماراجی استفاده کرد و به دقت شناسایی ۹۴/۶۶ درصد دست یافتند.

استوار (Ostovar) و همکاران [۱۸] یک مفهوم سطح بالاتر به عنوان «اجرا» تعریف کردند، این مفهوم به معنای مجموعه از پیمایش‌ها است که در اجرای فرآیند کسب و کار، با هم به صورت موازی اجرا شوند. استوار و همکاران علاوه بر اطلاعات آماری «اجرا» در سابقه‌ی رویدادها از روابط دوتایی بین رویدادها که در الگوریتم آلفا+ معرفی شده بود برای شناسایی رانش مفهوم استفاده کردند. این روابط دوتایی شامل هم روندی[۱۳] تقدم-تاخر[۱۴] حلقه به خود[۱۵]و حلقه به طول دو[۱۶] (دوتایی) می‌باشد. استوار و همکاران در ادامه‌ی تحقیقات خود [۱۳] با معرفی درخت فرآیند و استخراج درخت فرآیند در دو پنجره‌ی ارجاع و تشخیص به جای استخراج مدل فرآیند، دقت و سرعت روش خود را بهبود دادند که باعث شد به دقت ۹۹/۷۴ درصدی دست یابند.

آکورسی و همکاران [۱۹] و هومپز و همکاران [۲۰] هم از روش‌های متفاوتی برای شناسایی رانش مفهوم استفاده می‌شود. آنها از خوشه‌بندی پیمایش‌ها و محاسبه‌ی فاصله‌ی هر جفت رویداد در هر پیمایش استفاده کردند. در حقیقت ساختار فرآیند در قالب محل قرار گرفتن هر رویداد در پیمایش در روش آنها مورد توجه قرار گرفت. در روش آنها هم مانند روش بوز و همکاران اندازه‌ی پنجره باید توسط کاربر تنظیم می‌شد. کارمونا و همکاران [۱۲] یک روش بلادرنگ با توانایی آموزش را برای شناسایی رانش مفهوم ارائه شد. در روش آنها اندازه‌ی پنجره طبق یک الگوریتم یادگیری تخمین زده می‌شد سپس یک پنجره‌ی انطباقی با استفاده از پارامتر بدست آمده نقاط رانش مفهوم را شناسایی می‌کرد.

۳- روش پژوهش

سیستم‌های مدیریت فرآیندهای کسب و کار (BPMS) حداقل شامل سه بخش اساسی هستند. یک بخش شامل یک ویرایش-گر گرافیکی قدرتمند برای طراحی مدل فرآیند است و بخش دوم که می‌توان مهم‌ترین بخش این سیستم‌ها دانست موتور اجرای فرآیند سازمانی است و بخش سوم آن تحلیل و بهینه سازی اجرای فرآیندها می‌باشد. در سازمان‌هایی که از BPMS برای مدیریت و اداره‌ی سیستم‌های اطلاعاتی خود استفاده می-کنند، هر فرآیند برای پاسخ دهی به یک درخواست مشتری یا یک درخواست از سایر بخش‌های سازمان طراحی می‌شود. با ایجاد یک درخواست، از مدل فرآیند مربوطه با آن درخواست یک نمونه فرآیند[۱۷]ایجاد می‌شود و این نمونه فرآیند در موتور اجرای فرآیند اجرا می‌شود. در جریان اجرای یک فرآیند مجموعه‌ای از رویدادها رخ می‌دهد که مهم‌ترین آنها ایجاد یک

---

[۱] Traces
[۲] Concurrency
[۳] Causality
[۴] Self-loop
[۵] Length-two loop
[۶] Process Instance

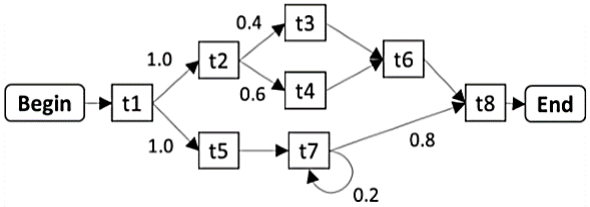
کار، تخصیص یک کار به منبع(نیروی انسانی)، شروع انجام کار و اتمام کار می‌باشد و دنباله ای از این رویدادها به صورت یک یا چند فایل که غالباً به دو فرمت XES یا MXML است به عنوان سابقه‌ی رویداد ذخیره می‌شوند. تحلیل سابقه‌ی ایجاد شده از اجرای فرآیندها پیشین می‌تواند باعث بهبود و بهینه سازی اجرای فرآیندهای آتی شود. یکی از این موارد می‌تواند شناسایی رانش مفهوم در سابقه‌ی رویداد باشد. در ادامه قبل از بیان ایده‌ی اصلی این مقاله لازم است برخی از مفاهیم موجود در این زمینه را تعریف کنیم.

۳-۱ مفاهیم پایه و تعاریف

۳-۱-۱ مدل فرآیند، سابقه رویداد، پیمایش، گونه

مدل فرآیند $M = (T.C.F.start.end)$ پنج‌تایی است که شامل مجموعه‌ای از کارها[۱] $T = \{t_1.t_2.t_3.\dots t_n\}$ است و مجموعه‌ی مشخصی از راس‌های کنترلی $C = \{AND_{join}.AND_{split}.XOR_{join}.XOR_{split}\}$ و دو راس $start$ و end به عنوان راس‌های شروع و پایان است که در یک گراف جهت‌دار $F = (V.E)$ به هم وصل شده‌اند به‌طوری‌که $V = T \cup C \cup \{start.end\}$ و $E = V \times V$ که بیانگر تقدم و تاخر انجام کارها در مدل فرآیند است.

یک سابقه‌ی رویداد $\mathcal{L}$ شامل مجموعه‌ای از پیمایش‌ها $\mathcal{L} = \{\tau_1.\tau_2.\tau_3.\dots \tau_L\}$ است. که $L = |\mathcal{L}|$ اندازه‌ی سابقه‌ی رویداد یا همان تعداد پیمایش‌های موجود در سابقه‌ی رویداد را مشخص می‌کند. هر پیمایش خود شامل دنباله‌ای از رویدادهایی است که در اجرای یک نمونه فرآیند مربوط به یک درخواست مشخص در سابقه‌ی رویداد ثبت می‌شود. بنابراین هر پیمایش را می‌توان به صورت $\tau = \langle e_1^a.e_2^b.e_3^c.\dots.e_k^d \rangle$ نشان داد. در هر عنصر e_i^a از این پیمایش متناظر با رویدادی که در زمان i ام روی کار $a \in T$ از فرآیند رخ داده است. از آنجایی که در گراف F می‌تواند حلقه وجود داشته باشد؛ مقدار a در رویدادهای یک پیمایش می‌تواند تکراری باشد. تابع $\rho: \tau \to T^*$ را به عنوان مسیر حرکت پیمایش τ روی گرافF تعریف می‌کنیم. پیمایش-های موجود در سابقه‌ی رویداد را می‌توان بر اساس مسیر حرکتش روی گراف F دسته بندی کرد به عبارت دیگر دو پیمایش $\tau_i.\tau_j \in \mathcal{L}$ که $\rho(\tau_2) = \rho(\tau_1)$ یک مسیر یکسان را روی گراف F پیمایش می‌کنند؛ در یک دسته قرار می‌گیرند در ادامه این مقاله به هر دسته از پیمایش‌ها یک «*گونه*»[۱] می-گوییم. به عنوان مثال در شکل(۳) اجرای فرآیند تولید [۲۱] به دلیل وجود حلقه روی کار t_7 می‌تواند به صورت تئوری بینهایت گونه تولید کند.



گونه ۱	$\langle t_1.t_2.t_3.t_5.t_6.t_7.t_8 \rangle$
گونه ۲	$\langle t_1.t_2.t_5.t_7.t_4.t_6.t_8 \rangle$
گونه ۳	$\langle t_1.t_2.t_3.t_5.t_6.t_7.t_7.t_8 \rangle$
گونه ۴	$\langle t_1.t_2.t_5.t_3.t_6.t_7.t_8 \rangle$
⋮	⋮
گونه n	$\langle t_1.t_2.t_5.t_3.t_7.t_6.t_7.t_7.t_8 \rangle$

شکل ۳. مدل فرآیند تولید [21]

۳-۱- ایده‌ی اصلی

از آنجایی که عامل‌های اصلی رانش مفهوم به نوع تغییر در فرآیند برمی‌گردد. هر نوع تغییری در فرآیند کسب و کار (به جز تغییر در احتمال انتخاب انشعاب XOR) در طول(تعداد رویدادهای موجود در پیمایش) پیمایش و دسته‌بندی گونه‌ی پیمایش اثر گذار خواهد بود. و باعث تغییر در توزیع و فراوانی گونه‌ها می‌شود. در این مقاله برای شناسایی رانش مفهوم از دو پنجره‌ی زمانی با عنوان‌های پنجره‌ی «ارجاع» و پنجره‌ی «تشخیص» استفاده می‌شود. این دو پنجره در طول زمان روی سابقه‌ی رویداد حرکت می‌کنند. طول هر دو پنجره با هم برابر است و عبارت است از تعداد پیمایش‌هایی از سابقه‌ی رویداد که در این پنجره‌ها قرار می‌گیرد. ایده‌ی اصلی در شناسایی رانش مفهوم، تغییر توزیع آماری گونه‌های موجود در این دو پنجره‌ی ارجاع و تشخیص می‌باشد.

۳-۲- شناسایی رانش مفهوم براساس فراوانی گونه

روند کلی روش پیشنهادی در شکل(۴) آمده است. در ابتدا، دو پنجره با اندازه‌ی یکسان روی سابقه‌ی رویداد حرکت می‌کند از هر پنجره یک بردار از میزان فراوانی جفت کارهای پشت سر هم داخل پنجره ایجاد می‌شود. در این مرحله دو بردار فراوانی یا بردار ویژگی به نام‌های V_R و V_D ایجاد می‌شود سپس با یک آزمون آماری یا یک تابع مشخص، فاصله‌ی این دو بردار محاسبه می‌شود. با حرکت دو پنجره بر روی کل سابقه‌ی رویداد و محاسبه‌ی فاصله‌ی بردارهای ویژگی، نمودار تغییرات فاصله برای کل سابقه‌ی رویداد ایجاد می‌شود. سپس در مرحله‌ی بعد

[۱] Task

[۱] Variant

با تعیین مقدار آستانه، نقاط رانش مفهوم شناسایی می‌شود. بنابراین با توجه به شکل(۴) می‌توان مراحل شناسایی رانش مفهوم را در روش پیشنهادی به صورت ذیل بیان کرد.

۱) ایجاد دو پنجره با اندازه‌ی w تحت عنوان پنجره‌ی ارجاع $W_R = \{\tau_{i-w}.\tau_{i-w+1}.\tau_{i-w+2}.\dots\tau_i\}$ و پنجره‌ی تشخیص $W_D = \{\tau_{i+1}.\tau_{i+2}.\tau_{i+3}.\dots\tau_{i+w+1}\}$ بر روی سابقه‌ی رویداد به‌طوری‌که در مجموع $2w$ پیمایش از سابقه‌ی رویداد را در بر بگیرد.

۲) استخراج بردار ویژگی برای هر پنجره با شمارش جفت کارهای پشت سر هم در گونه‌های موجود در هر پنجره به نام‌های V_R و V_D به ترتیب برای پنجره‌های ارجاع و تشخیص

۳) محاسبه‌ی فاصله دو بردار ویژگی بر اساس رابطه(۴) در بخش ۳-۳-۱ (تابع پیشنهادی این مقاله)

۴) تکرار مراحل ۱ الی ۳ با حرکت دادن دو پنجره در طول زمان تا کل پیمایش‌های موجود در سابقه‌ی رویداد مشاهده شوند و استخراج تغییرات فاصله بردارهای V_R و V_D

۵) محاسبه‌ی مقدار آستانه با استفاده از هیستوگرام مقادیر فاصله‌ی بردارهای ویژگی

۳-۲-۱- محاسبه‌ی فاصله‌ی دو بردار ویژگی

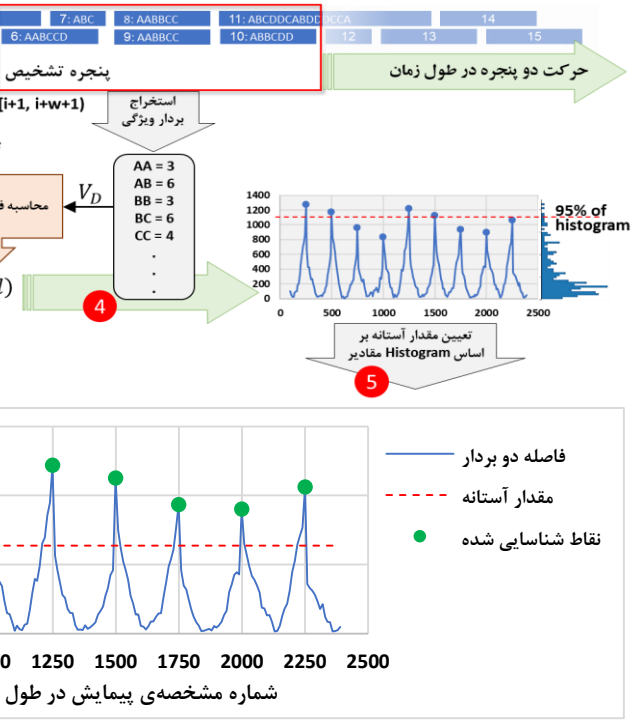
در این مقاله برای محاسبه‌ی فاصله‌ی دو بردار ویژگی از دو آزمون آماری χ^2 و $gTest$ و فاصله‌ی زاویه‌ای دو بردار (کسینوسی) و یک تابع فاصله‌ی پیشنهادی استفاده کرده‌ایم که نتایج آن به صورت تفکیک شده در بخش نتایج گزارش می‌شود. از آنجایی که در تابع فاصله‌ی پیشنهادی مشابه به آزمون آماری $gTest$ است. آنرا در ادامه $symgTest$ می‌نامیم.

$$chi2\_Test = \chi^2 = \sum_{i=1}^{n} \frac{(V_i^R - V_i^D)^2}{V_i^D} \quad (۱)$$

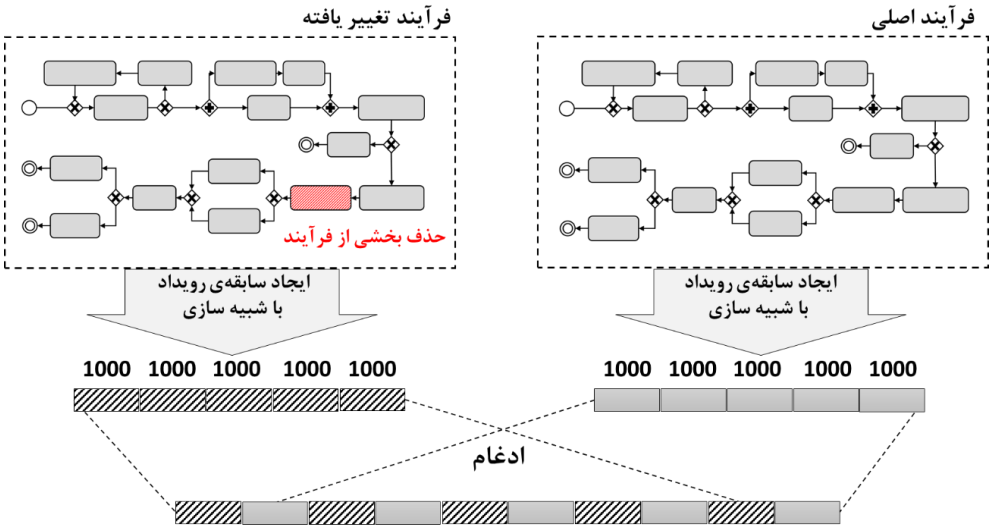
$$gTest = 2\sum_{i=1}^{n} V_i^D . ln\left(\frac{V_i^D}{V_i^R}\right) \quad (۲)$$

$$\cos(\theta) = \frac{V^R \cdot V^D}{\|V^R\| \times \|V^D\|} = \frac{\sum_{i=1}^{n} V_i^D V_i^R}{\sqrt{\sum_{i=1}^{n}(V_i^D)^2}\sqrt{\sum_{i=1}^{n}(V_i^R)^2}} \quad (۳)$$

$$symgTest = \sum_{i=1}^{n} |V_i^R - V_i^D| \times \left| ln\left(\frac{V_i^R}{V_i^D + \varepsilon}\right)\right| \quad (۴)$$



شکل ۴. روند کلی روش پیشن



شکل ۵. روش ایجاد فایلهای پایگاه داده شناسایی رانش مفهوم [۲۲]

در رابطه‌های (۱) و (۲) هر پیمایش به عنوان یک متغیر تصادفی در نظر گرفته می‌شود و جامعه آماری در پنجره‌ی ارجاع را به عنوان مرجع قرار می‌دهد و انتظار دارد نمونه‌های(پیمایش-های) مشاهده شده در پنجره تشخیص جامعه‌ی آماری مشابهی با پنجره‌ی ارجاع داشته باشد. رابطه‌ی(۳) فاصله کسینوسی دو بردار را نشان می‌دهد و در مواقعی که فضای بردارهای ویژگی فضای متریک است و اگر تغییرات فاصله‌ی دو بردار به صورت نسبی مشابه هم باشد، می‌تواند نتایج بهتری را در شناسایی رانش مفهوم نشان دهد. رابطه‌ی(۴) نسخه‌ی متقارن شده‌ی رابطه‌ی(۲) است که در این مقاله پیشنهاد شده است، این تغییر توانسته علاوه بر متقارن کردن فاصله $gTest$ با اضافه کردن پارامتر ε نقش پیمایش‌هایی که در یکی از دو پنجره وجود دارد و در دیگری وجود ندارد را برجسته کند. هرچه مقدار ε کوچک-تر باشد این اثر برجسته‌تر می‌شود.

۴- تجزیه و تحلیل یافته‌ها

ماراجی و همکاران [۷] برای آزمایش روش پیشنهادی خود ۷۲ فایل سابقه‌ی رویداد در فرمت MXML با استفاده از شبیه‌ساز BIMP[۲۰] ایجاد کردند. ماراجی برای ایجاد پایگاه دادگان خود، روی فرآیند درخواست وام (شکل(۲)) تغییرات جدول(۱) را ایجاد کرد و هر دو فرآیند اصلی و تغییر یافته را در شبیه ساز BIMP برای تعداد درخواست ۵۰۰۰، ۳۷۵۰، ۲۵۰۰ و ۱۲۵۰ اجرا شده است و مانند شکل (۵) درخواست به ۵ دسته‌ی به ترتیب ۱۰۰۰ یا ۷۵۰، ۵۰۰ و ۲۵۰ تایی تقسیم شده و به صورت یکی در میان از فرآیند اصلی و تغییر یافته با هم ادغام شدند و در نهایت به ترتیب فایل‌هایی شامل ۱۰۰۰۰ یا ۷۵۰۰، ۵۰۰۰ و ۲۵۰۰ پیمایش ایجاد شد. هر فایل شامل ۹ رانش مفهوم از فرآیند اصلی به تغییر یافته و بعکس را دارا است. به همین ترتیب برای هر یک از ۱۲ عامل لیست شده در جدول(۱) (و ۶ عامل با ترکیب و انتخاب یک عامل از هر گروه) چهار فایل دادگان ایجاد کردند. به عنوان مثال برای «حذف یا ایجاد یک بخش در فرآیند» چهار فایل به عنوان re10k، re7.5k، re5k و re2.5k با فرمت MXML ایجاد شده است.

۴-۱- روش اندازه گیری دقت شناسایی

برای اندازه گیری دقت شناسایی از معیار F_1 استفاده می‌کنیم که روش محاسبه‌ی آن در رابطه‌ی (۷) آمده است.

$$precision = \frac{True\ positives}{True\ positives + False\ positives} \quad (۵)$$

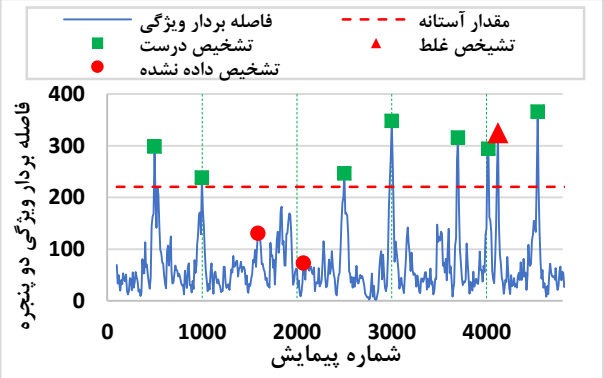
$$recall = \frac{True\ positives}{True\ positives + False\ negatives} \quad (۶)$$

$$F_1 = 2 \cdot \frac{precision \cdot recall}{precision + recall} \quad (۷)$$

در هر فایل از سابقه‌ی رویداد ۹ رانش مفهوم وجود دارد. در شکل ۶) یک نمونه از اجرای روش پیشنهادی روی فایل fr5k را نشان می‌دهد در این فایل ۹ رانش در محل‌های $500k$ برای مقادیر $k = 1..9 \in \mathbb{N}$ وجود دارد. در این اجرا ۷ رانش مفهوم به درستی تشخیص داده شده است، یک رانش مفهوم در شماره پیمایش ۴۱۱۹ به اشتباه تشخیص داده شده است و دو رانش

[۲۰] http://bimp.cs.ut.ee/

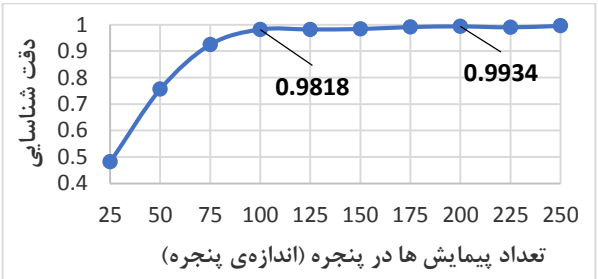
مفهوم در محلهای ۱۵۰۰ و ۲۰۰۰ تشخیص داده نشده است. بنابراین طبق رابطه‌ی (۷) مقدار $F_1 = 2\frac{0.875\times0.778}{0.875+0.778} = 0.823$ خواهد شد. به همین صورت مقدار F_1 برای۷۲ اجرا محاسبه خواهد شد.



شکل ۶- نمودار فاصله بردار ویژگی بین دو پنجره‌ی ارجاع و تشخیص حاصل از اجرای الگوریتم پیشنهادی روی فایل fr5k

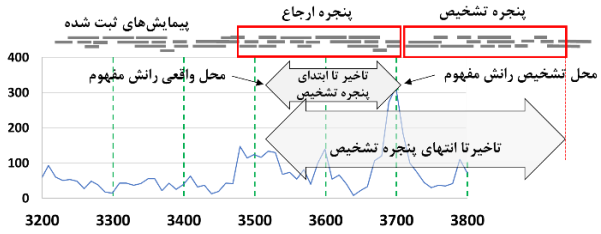
۴-۲- اثر اندازه پنجره در دقت شناسایی

از آنجایی که الگوریتم پیشنهادی بر اساس حرکت دو پنجره‌ی ارجاع و تشخیص طراحی شده است اندازه‌ی این دو پنجره در دقت شناسایی تاثیر گذار است. هرچه اندازه‌ی پنجره بزرگتر باشد پیمایش‌های بیشتری در هر پنجره بررسی می‌شود و بردارهای ایجاد شده ابعاد بزرگتری خواهد داشت، در نتیجه باعث بهبود دقت شناسایی می‌شود. برای نمایش این اثر در روش پیشنهادی از تابع فاصله $symgTest$ (رابطه‌ی(۴)) و مقادیر اندازه‌ی پنجره ۲۵ الی ۲۵۰ پیمایش با افزایش ۲۵تایی استفاده می‌شود و مقدار میانگین آن برای ۷۲ فایل ورودی پایگاه دادگان محاسبه می‌شود که نتیجه‌ی آن در شکل (۷) نمایش داده شده است.

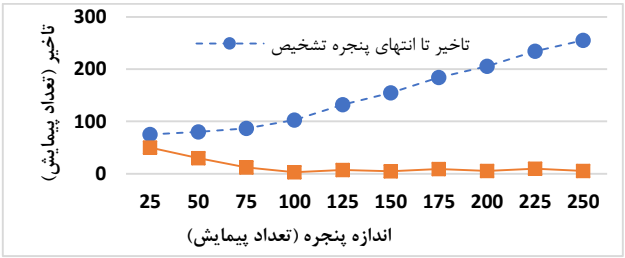


شکل ۷- اثر اندازه‌ی پنجره های تشخیص و ارجاع در دقت شناسایی (F_1) رانش مفهوم

در شناسایی رانش مفهوم علاوه بر دقت شناسایی زمان یا سرعت تشخیص رانش نیز اهمیت دارد، مخصوصاً در کاربردهای برخط که بعد از شناسایی رانش توسط مدیریت تصمیماتی در مورد عملکرد سیستم گرفته می‌شود، بنابراین شناسایی به موقع و زود هنگام رانش مفهوم از اهمیت ویژه‌ای در این حیطه برخوردار است. همانطور که در شکل ۷) مشاهده شد؛ افزایش اندازه‌ی پنجره باعث بهبود در دقت شناسایی رانش مفهوم شد. اما همین افزایش می‌تواند باعث افزایش تاخیر در شناسایی شود. برای این منظور در آزمایش انجام شده مطابق با شکل ۸) برای شناسایی-های درست تشخیص داده شده میانگین تاخیر به ازای هر اندازه پنجره اندازه‌گیری شد که نتیجه‌ی آن در شکل ۹) آمده است.

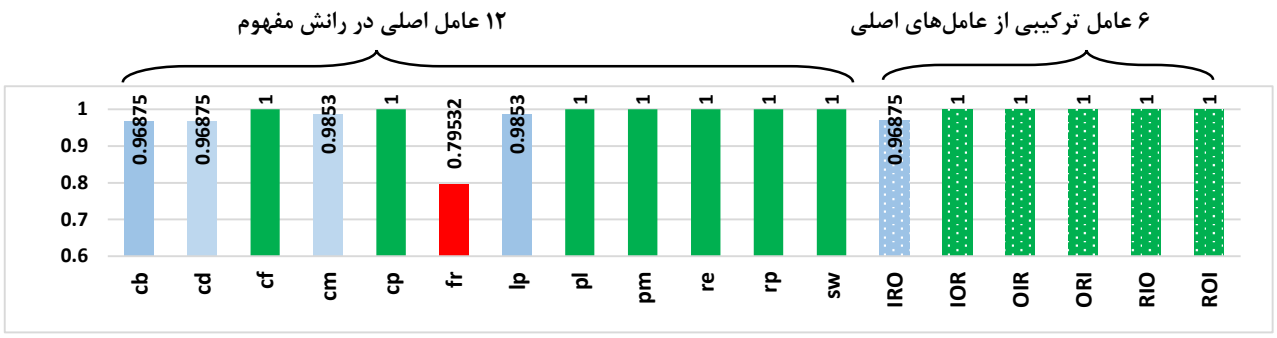


شکل ۸- روش محاسبه‌ی تاخیر در شناسایی رانش مفهوم

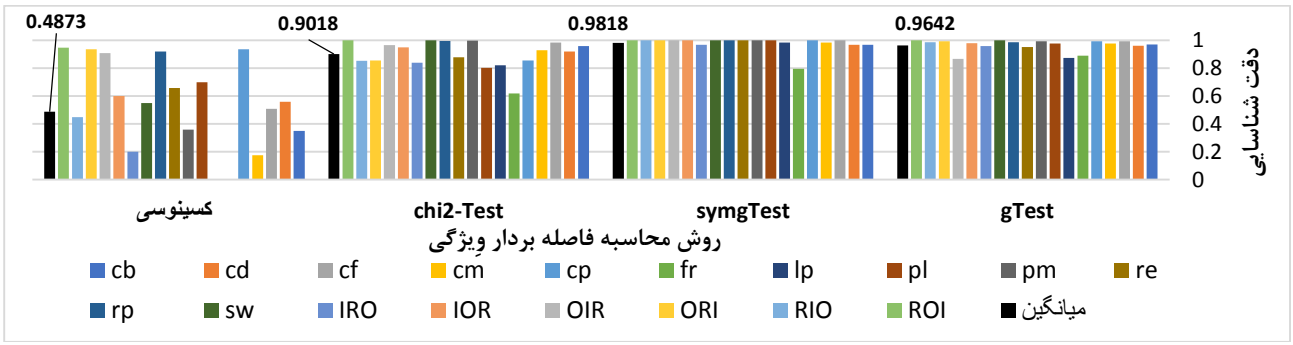


شکل ۹- اثر اندازه پنجره در تاخیر شناسایی رانش مفهوم

همان‌طور که در شکل ۹) نشان داده شده است، افزایش اندازه‌ی پنجره‌ی تاخیر شناسایی را تا ابتدای پنجره‌ی تشخیص کاهش می‌دهد ولی از آنجایی که در پیشینه‌ی پژوهش تاخیر شناسایی تا انتهای پنجره تشخیص تعریف شده است. متاسفانه با افزایش اندازه‌ی پنجره تاخیر افزایش می‌یابد. این اثر در شکل ۹) در نمودار تاخیر تا انتهای پنجره‌ی تشخیص مشهود است. بنابراین با اینکه افزایش اندازه‌ی پنجره باعث بهبود دقت شناسایی می-شود ولی در کاربردهای برخط که کاهش تاخیر شناسایی نیز دارای اهمیت است باید اندازه پنجره خیلی بزرگ نباشد. در نتیجه با توجه به نمودارهای دقت شناسایی و تاخیر شناسایی که در شکل ۷) و شکل ۹) آمده است می‌توان اندازه‌ی پنجره‌ی ۱۰۰ را مقدار مناسبی که نتیجه‌ی این دو تقابل است در نظر گرفت. که هم دقت به اندازه‌ی کافی افزایش یافته است و هم تاخیر هنوز خیلی زیاد نشده است. البته نگاه دقیق‌تر به پیشینه-ی پژوهش، اندازه‌ی پنجره‌ی ۱۰۰ را روی پایگاه دادِگان مورد آزمایش تایید می‌کند.



شکل ۱۰- دقت شناسایی به تفکیک عامل‌های رانش



شکل ۱۱- مقایسه توابع فاصله بردارهای ویژگی در میزان دقت شناسایی

۴-۳- اثر عامل‌های رانش مفهوم در دقت شناسایی

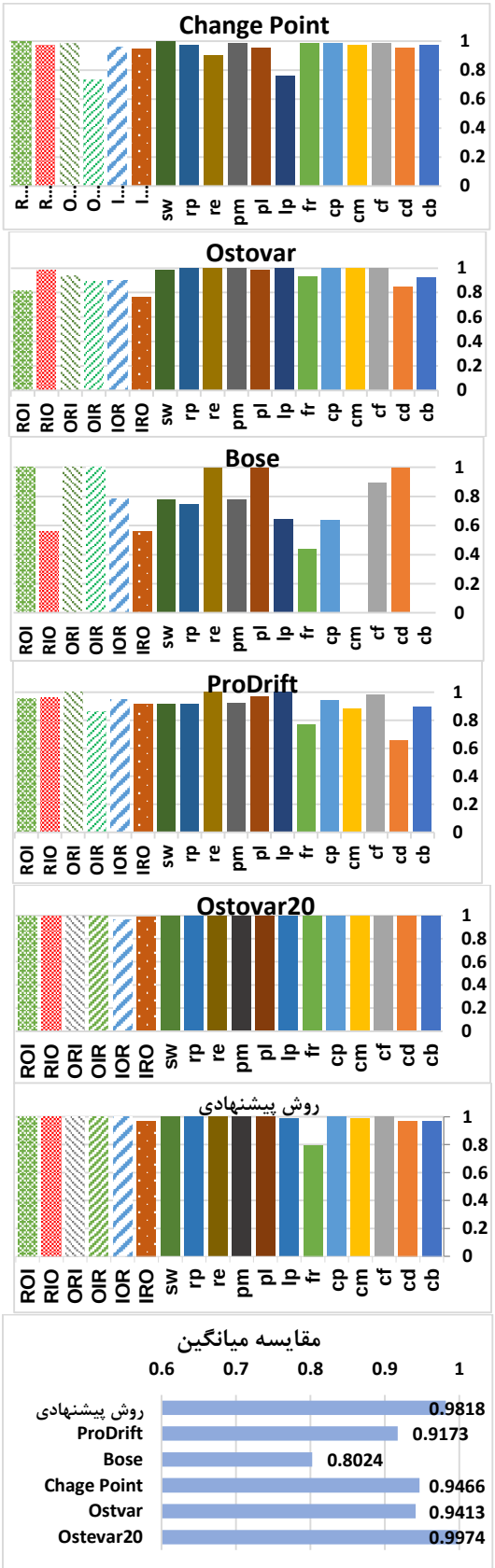
برای آزمایش دقیق‌تر اثر عامل‌های رانش بر روی روش پیشنهادی، با اینکه در پایگاه دادگان برای ۱۲ ماهیت رانش (و ۶ نوع ترکیبی) ۴ فایل وجود داشت و در هر فایل ۹ رانش با فاصله های ۲۵۰، ۵۰۰، ۷۵۰ و ۱۰۰۰ پیمایش؛ برای حذف اثر اندازه‌ی پنجره در نتایج آزمایش برای مقادیر ۲۵ الی ۲۵۰ با فاصله ۲۵تایی (در مجموع ۱۰ اندازه پنجره) تکرار شد. در واقع برای هر نوع رانش ۴۰ آزمایش انجام شد. میانگین دقت شناسایی در شکل ۱۰) نشان داده شده است. همانطورکه در شکل ۱۰) مشهود است بجز رانش مفهوم از نوع fr (تغییر در فراوانی انشعاب انحصاری) در سایر عامل‌ها دقت بالای ۹۶ درصد بدست آمده است و در ۱۲ مورد به دقت ۱۰۰ درصد حاصل شده است. با توجه به نتایج حاصل می‌توان به این نتیجه رسید که رانش‌های مفهومی که در سابقه‌ی رویداد باعث ایجاد و یا حذف یک گونه از پیمایش‌ها می‌شوند روش پیشنهادی در شناسایی آنها دقت بالایی دارد و در مواردی که فقط فراوانی گونه‌های تغییر می‌کند دقت کمتری خواهد داشت.

۴-۴- اثر نوع توابع فاصله در دقت شناسایی

در این مقاله علاوه بر تابع symgTest (تابع پیشنهادی) به عنوان فاصله‌ی دو بردار ویژگی در پنجره‌های ارجاع و تشخیص سه تابع gTest ، Chi2-Test و فاصله کسینوسی (رابطه‌های (۱) الی (۳)) مورد آزمایش قرار گرفتند. شکل (۱۱) اثر بهتر تابع پیشنهادی را نسبت به سایر توابع را نشان می‌دهد. لازم به ذکر است تابع gTest، تابع استفاده شده در سایر پژوهش‌ها ذکر شده در این مقاله است. مقایسه با الگوریتم های موجود

۴-۴-۱- از نظر دقت شناسایی

جهت مقایسه‌ی روش پیشنهادی، چهار روش از بهترین الگوریتم‌های موجود در پیشینه‌ی پژوهش، استوار [۱۳،۱۸]، ProDrift [۲۲] و Change Point [۱۷] و یک الگوریتم پایه‌ای که توسط بوز (Bose) و همکاران[۱۶] ارائه شده است را انتخاب کردیم. دو تا روش‌های مورد مقایسه، از استوار و همکاران انتخاب شده است که یکی از آنها در ۲۰۱۶ ارائه شده است و دیگری در مارس ۲۰۲۰ ارائه شده است که بهترین الگوریتم ارائه شده در شناسایی رانش مفهوم تا کنون است. برای مشخص شدن این دو روش از یکدیگر، در ادامه از عناوین Ostovar و Ostovar20 استفاده شده است. نتایج اجرای الگوریتم پیشنهادی و پنج الگوریتم انتخاب شده بر روی ۷۲ فایل پایگاه داده معرفی شده توسط ماراجی [۷]، در شکل (۱۲) نمایش داده شده است. در این آزمایش اندازه‌ی پنجره‌ها برای هر الگوریتم، مقدار(مقدار اولیه برای پنجره انطباقی) ۱۰۰ پیمایش در نظر گرفته شده است، البته روش‌های Ostovar و Chage Point از پنجره های انطباقی(طول متغییر) استفاده کرده اند. نتایج به تفکیک ۱۲ عامل رانش مفهوم و ۶ عامل ترکیبی نمایش داده شده است.



شکل ۱۲- مقایسه‌ی روش‌های [13]Ostovar20، [18]Ostovar، [17]Change Point و [22]ProDrift با روش پیشنهادی با اندازه‌ی ۱۰۰ پیمایش در پنجره‌ی ارجاع و تشخیص

۴-۴-۲-از نظر سرعت

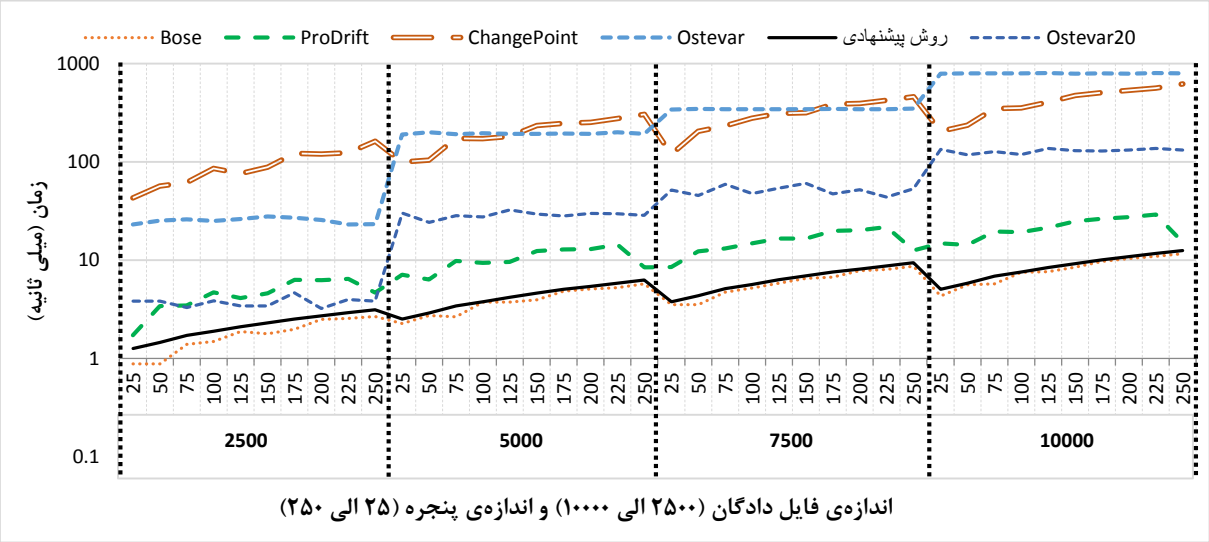
الگوریتم ارائه شده در این مقاله مشابه الگوریتم‌های Bose و ProDrift است با این تفاوت که نحوه‌ی استخراج بردارهای ویژگی از پنجره‌های ارجاع و تشخیص متفاوت است. در این الگوریتم‌ها دو پنجره با طول ثابت w پیمایش بر روی فایل دادگان با $|\mathcal{L}|$ پیمایش حرکت داده می‌شود و ویژگی های آماری با پیچیدگی زمانی $O(w)$ از هر پنجره استخراج می‌شود که در بدترین حالت پیچیدگی بررسی کل فایل دادگان $O(2w|\mathcal{L}|)$ خواهد شد. در روش پیشنهادی در هر پنجره به ازای هر پیمایش، جفت کارهای کنار هم نیز شمارش می‌شود که باعث می‌شود ضریب t^2 در پیچیدگی زمانی آن ظاهر می‌شود. در الگوریتم Change Point نیز از دو پنجره با طول متوسط w استفاده می‌شود (طول پنجره در زمان اجرا تغییر می‌کند) با این تفاوت که در هر پنجره به جای استخراج ویژگی‌های آماری، مدل فرآیند هر پنجره با استفاده از الگوریتم فرآیندکاوی ابتکاری(HM) استخراج می‌شود سپس از روی گراف حاصل شده ویژگی‌های متریک آن استخراج می‌گردد که در مجموع در بدترین حالت استخراج مدل فرآیند و ویژگی‌ها آن $O(w+t^2+wl)$ زمان خواهد برد که w ، t و l به ترتیب متوسط طول پنجره‌ی تطبیقی، تعداد کارهای(Task) مدل فرآیند (تعداد راس‌های گراف) و متوسط طول پیمایش است. الگوریتم‌های Ostovar و همکاران هم بر اساس دو پنجره با اندازه‌ی تطبیقی شناسایی رانش مفهوم را در بدترین حالت با پیچیدگی زمانی $O(|\mathcal{L}|^2)$ انجام می‌دهد که مستقل از اندازه‌ی پنجره است[۱۸].

جدول ۲- پیچیدگی زمان الگوریتم‌های مقایسه شده

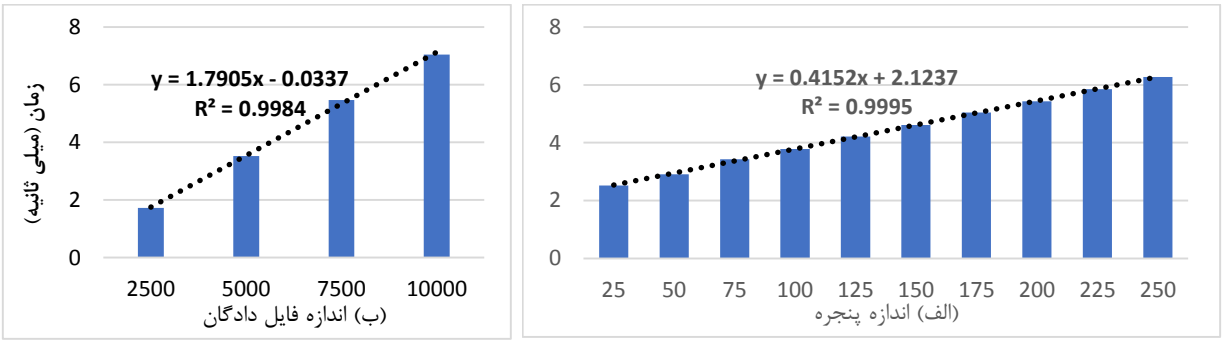
الگوریتم	پیچیدگی زمانی
Ostovar	$O(\|\mathcal{L}\|^2)$
Ostovar۲۰	$O(\|\mathcal{L}\|^2)$
Change Point	$O(2(w+t^2+wl)\|\mathcal{L}\|)$
ProDrift	$O(2w\|\mathcal{L}\|))$
Bose	$O(2w\|\mathcal{L}\|)$
روش پیشنهادی	$\boldsymbol{O(2wt^2\|\mathcal{L}\|)}$

برای آزمایش پیچیدگی زمانی تحلیل شده برای الگوریتم پیشنهادی، زمان اجرای الگوریتم پیشنهادی روی یک پردازنده Core i3 8109U با فرکانس 3GHz با مقدار حافظه 4GB در زبان برنامه‌نویسی Python اندازه‌گیری شده است. در شکل(۱۳) مقایسه‌ی زمان اجرای روش پیشنهادی با سایر االگوریتم‌ها را

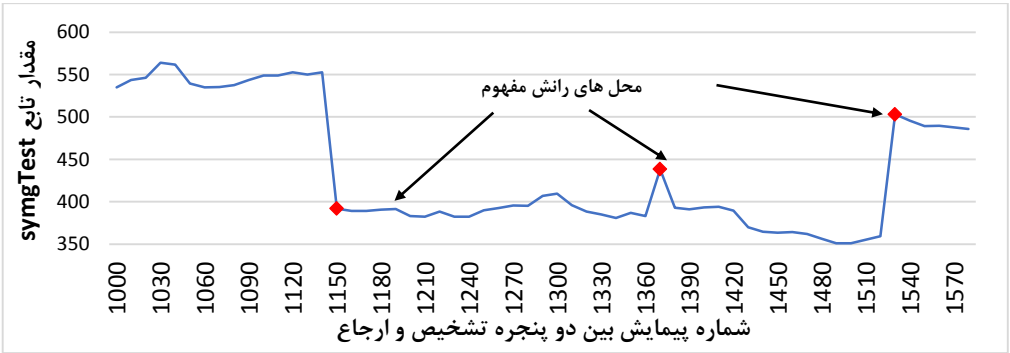
نشان می‌دهد. این آزمایش روی ۷۲ فایل (۱۸ الگو در چهار اندازه فایل دادگان) و اندازه پنجره‌ی ۲۵ الی ۲۵۰ پیمایش انجام شده است. همانطور که در جدول ۲ و شکل (۱۳) مشخص است سه الگوریتم Bose، ProDrift و روش پیشنهادی زمان اجرای خیلی کمتری نسبت به سه الگوریتم Ostovar، Ostovar20 و ChangePoint هست که می‌توان علت اصلی این تفاوت را در استفاده از پنجره انطباقی دانست، درحالی که دقت شناسایی الگوریتم پیشنهادی از بهترین روش کمتر از یک درصد است. در شکل(۱۳) نتایج اجرای روش پیشنهادی به تفکیک اندازه‌ی پنجره و اندازه‌ی فایل دادگان به صورت جداگانه نمایش داده شده است. همانطورکه در شکل(۱۴) مشخص است زمان اجرای الگوریتم پیشنهادی، با افزایش اندازه‌ی پنجره و افزایش فایل دادگان به صورت خطی افزایش یافته است.



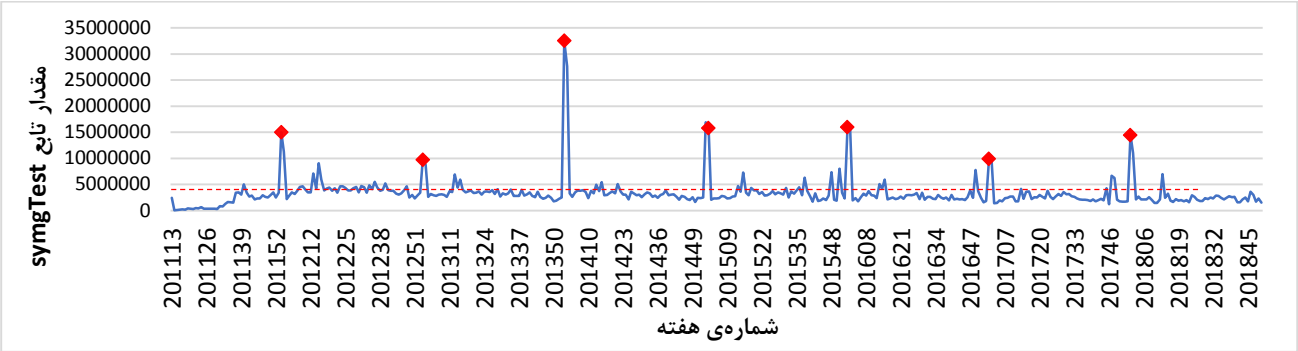
شکل ۱۳- اندازه‌گیری زمان اجرا و مقایسه‌ی روش پیشنهادی با سایر روش‌ها (محور افقی تغییرات اندازه‌ی فایل دادگان و اندازه‌ی پنجره را نشان می‌دهد و محور عمودی زمان اجرای الگوریتم را به میلی ثانیه در مقیاس لگاریتمی نشان می‌دهد)



شکل ۱۴- اندازه‌گیری زمان اجرای الگوریتم پیشنهادی (الف) برای مقادیر مختلف اندازه‌ی پنجره و (ب) برای مقادیر مختلف اندازه فایل دادگان که با افزایش آنها زمان استخراج رانش به صورت خطی افزایش میابد



شکل۱۵- استخراج رانش مفهوم روی پایگاه دادگان فرآیند مدیریت بلیط که دو رانش ناگهانی و یک رانش مقطعی را نشان می‌دهد



شکل۱۶- استخراج رانش مفهوم روی داده‌های ترافیکی شهر مقدس مشهد که هفت رانش مقطعی را نشان می‌دهد

۵- آزمایش روش پیشنهادی روی پایگاه دادگان واقعی

۵-۱- شناسایی رانش در فرآیند مدیریت بلیط

مشابه با آزمایش انجام شده توسط استوار و همکاران [13] روش پیشنهادی روی یک پایگاه دادگان فرآیند مدیریت بلیط[۱]در یک شرکت نرم افزاری ایتالیایی اجرا شد. در این پایگاه ۲۱۳۴۸ رویداد، ۱۴ فعالیت، ۴۵۸۰ پیمایش وجود دارد. در اجرای الگوریتم پیشنهادی مشابه آزمایش استوار، اندازه پنجره‌ها ۱۰۰۰ پیمایش در نظر گرفته شده است که با پرش‌های ۱۰ پیمایش در شکل (۱۵) نشان داده شده است. مشابه گزارش استوار و همکاران روش پیشنهادی در این مقاله نیز دو رانش اصلی در محل های ۱۱۵۰ و ۱۵۳۰ را نشان می‌دهد، روش پیشنهادی علاوه بر گزارش استوار و همکاران یک رانش کوچکتر را نیز در محل ۱۳۷۰ را یافته است که البته گزارش نشدن آن توسط استوار و همکاران می‌تواند به علت متفاوت بودن مقدار حد استانه تغییرات در شناسایی رانش مفهوم باشد.

۵-۲- شناسایی رانش روی داده‌های ترافیکی مشهد

پایگاه دادگان واقعی دومی که برای آزمایش در این مقاله استفاده شده است یک پایگاه داده غیر فرآیندی است که با یک تغییر جزئی در استخراج بردار ویژگی در روش پیشنهادی، امکان استخراج رانش‌های مفهوم در داده های ترافیک شهری را نشان می‌دهد و اهمیت موضوع مورد بحث در این مقاله را روشن‌تر می‌کند. پایگاه دادگان ترافیک شهری مشهد شامل ۵,۹۷۹,۰۷۳ رویداد ثبت اطلاعات استفاده از من‌کارت در ۲۶۹۶ ترمینال اتوبوسرانی و قطار شهری در شهر مقدس مشهد است که در بازه‌ی تاریخی ۲۰۱۱/۰۳ الی ۲۰۱۸/۰۳ ثبت شده است. با شمارش تعداد من‌کارت‌های استفاده شده در هر ترمینال بردار ویژگی هر پنجره حاصل می‌شود. در این آزمایش از پنجره‌های زمانی یک هفته‌ای با پرش‌های یک هفته استفاده شده است و از تابع پیشنهادی symgTest برای مقایسه‌ی بردارهای استفاده شده است. شکل (۱۶) محل‌های رانش‌های اصلی را در داده‌های ترافیکی شهر مقدس مشهد را نشان می‌دهد. در این شکل محل‌های رانش مفهوم مربوط به دو مناسبت رحلت رسول اکرم (ص) و شهادت امام رضا (ع) است که تاثیر چشم‌گیری در رفتار ترافیک شهری مشهد گذاشته است که با فواصل مرتب یکسال قمری در شکل (۱۶) مشهود است.

۶- جمع بندی و نتیجه گیری

در این مقاله یک روش ابتکاری جهت استخراج رانش مفهوم در فرآیندهای کسب و کار ارائه شد. در این روش با حرکت دو پنجره‌ی ارجاع و تشخیص، بر روی سابقه‌ی رویداد و استخراج مولفه‌های آماری پنجره‌ها و مقایسه‌ی آنها جهت شناسایی رانش استفاده شده است. ایده‌ی اصلی مقاله شمارش جفت کارهای انجام شده پشت سر هم در اجرای فرآیند است که ویژگی مفیدی در استخراج الگوهای رانش مفهوم است. علاوه بر این در این مقاله یک تابع مقایسه تحت عنوان symgTest معرفی شد که نسبت به gTest نتیجه‌ی بهتری (۱/۷۶ درصد) را در دقت شناسایی در روش پیشنهادی داشت. در صورتی که فاصله‌ی دو بردار از حد آستانه افزایش یابد، شناسایی رانش انجام می‌شود. برای محاسبه‌ی حد آستانه نیز از هیستوگرام نمودار فاصله‌ی دو بردار استفاده شده است. آزمایش عملکرد و دقت روش پیشنهادی بر روی یک پایگاه دادگان موجود در پیشینه‌ی پژوهش انجام شد. آزمایش انجام شده دقت شناسایی ۹۸/۱۸

[۱] https://doi.org/10.4121/uuid:0c60edf1-6f83-4e75-9367-4c63b3e9d5bb

درصد برای اندازه پنجره ۱۰۰ را نشان می‌دهد. این مقدار بهترین نتیجه‌ی حاصل در روشهای ارائه شده با اندازه پنجره ثابت است در حالی که بهترین دقت شناسایی (۹۹/۷۴ درصد) مربوط به استوار و همکاران روشی با اندازه پنجره تطبیقی است. تغییر اندازه پنجره در زمان شناسایی رانش مفهوم هزینه زمانی بیشتری را نسبت به روشها با پنجره ثابت دارد. با اینکه روش پیشنهادی دقت کمی کمتر از روش استوار دارد ولی هزینه زمانی بسیار کمتری را نسبت روش استوار دارد. در روش پیشنهادی هزینه زمانی نسبت به اندازه فایل دادگان به صورت خطی رشد می‌کند در حالی که روش استوار نسبت به اندازه فایل دادگان از نوع چندجمله‌ای درج دوم است. یکی از مزایای روش پیشنهادی عدم وابستگی روش به ساختار گراف در مدل فرآیند است و فقط از توالی جفت کارهای پشت سر هم استفاده می‌کند. به دلیل سادگی روش پیشنهادی در پیاده سازی، این روش را به راحتی می‌توان برای سایر داده‌های غیر فرآیندی نیز استفاده کرد.

مراجع

1. W. M. P. van der Aalst, *Process Mining: Discovery, Conformance and Enhancement of Business Processes*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2011.
2. R. P. J. C. Bose, W. M. P. van der Aalst, I. . Žliobait\.e, and M. Pechenizkiy, "Handling concept drift in process mining," in *International Conference on Advanced Information Systems Engineering*, 2011, pp. 391–405.
3. A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen, "Handling local concept drift with dynamic integration of classifiers: Domain of antibiotic resistance in nosocomial infections," in *19th IEEE Symposium on Computer-Based Medical Systems (CBMS'06)*, 2006, pp. 679–684.
4. M. Pechenizkiy, J. Bakker, I. Žliobaitė, A. Ivannikov, and T. Kärkkäinen, "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift," *ACM SIGKDD Explor. Newsl.*, vol. 11, no. 2, pp. 109–116, 2010.
5. M. Van Leeuwen and A. Siebes, "Streamkrimp: Detecting change in data streams," in *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 2008, pp. 672–687.
6. D. Brzezinski and J. Stefanowski, "Reacting to different types of concept drift: The accuracy updated ensemble algorithm," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 25, no. 1, pp. 81–94, 2014.
7. A. Maaradji, M. Dumas, M. La Rosa, and A. Ostovar, "Fast and accurate business process drift detection," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 9253, pp. 406–422, 2015.
8. M. Reichert, C. Hensinger, and P. Dadam, "Supporting adaptive workflows in advanced application environments," 1998.
9. S. Rinderle, M. Reichert, and P. Dadam, "Correctness criteria for dynamic changes in workflow systems, a survey," *Data Knowl. Eng.*, vol. 50, no. 1, pp. 9–34, 2004.
10. A. Adriansyah *et al.*, "Process mining manifesto," 2012.
11. R. Accorsi and T. Stocker, "Discovering workflow changes with time-based trace clustering," in *International Symposium on Data-Driven Process Discovery and Analysis*, 2011, pp. 154–168.
12. J. Carmona and R. Gavalda, "Online techniques for dealing with concept drift in process mining," in *International Symposium on Intelligent Data Analysis*, 2012, pp. 90–102.
13. A. Ostovar, S. J. J. Leemans, and M. La Rosa, "Robust drift characterization from event streams of business processes," *ACM Trans. Knowl. Discov. from Data*, vol. 14, no. 3, pp. 1–57, 2020.
14. J. Martjushev, R. P. J. C. Bose, and W. M. P. van der Aalst, "Change point detection and dealing with gradual and multi-order dynamics in process mining," in *International Conference on Business Informatics Research*, 2015, pp. 161–178.
15. C. W. Günther, S. Rinderle, M. Reichert, and W. Van Der Aalst, "Change mining in adaptive process management systems," in *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, 2006, pp. 309–326.
16. R. P. J. C. Bose, W. M. P. Van Der Aalst, I. . Žliobait\.e, and M. Pechenizkiy, "Dealing with concept drifts in process mining," *IEEE Trans. neural networks Learn. Syst.*, vol. 25, no. 1, pp. 154–171, 2014.
17. A. Seeliger, T. Nolle, and M. Mühlhäuser, "Detecting Concept Drift in Processes using Graph Metrics on Process Graphs," pp. 1–10, 2017.
18. A. Ostovar, M. Abderrahmane, M. La Rosa, A. H. ter Hofstede, and B. F. van Dongen., "Detecting drift from event streams of unpredictable business processes," in *International Conference on Conceptual Modeling*, 2016, pp. 330–346.
19. R. Accorsi and T. Stocker, "Discovering workflow changes with time-based trace clustering," in *International Symposium on Data-Driven Process Discovery and Analysis*, 2011, pp. 154–168.
20. B. Hompes, J. C. A. M. Buijs, W. M. P. van der Aalst, P. Dixit, and H. Buurman, "Detecting Change in Processes Using Comparative Trace Clustering.," in *SIMPDA*, 2015, pp. 95–108.
21. Y. Xie, C. F. Chien, and R. Z. Tang, "A dynamic task assignment approach based on individual worklists for minimizing the cycle time of business processes," *Comput. Ind. Eng.*, vol. 99, no. 12, pp. 401–414, 2016.
22. A. Maaradji, M. Dumas, M. La Rosa, and A. Ostovar, "Detecting sudden and gradual drifts in business processes from execution traces," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 10, pp. 2140–2154, 2017.

Fast and accurate concept drift detection from event logs

Abstract

In organizations and large companies that are using business process management systems (BPMSs), process model can change due to upstream laws, market conditions. BPMSs have flexible to these changes. Effect of these change are saved in storage devises and event logs; these changes are sometimes applied suddenly or gradually on the event logs. Changing the season or starting a new financial term can be a factor to make these changes. This change is called concept drift in business process model. On time detection and recognition of process concept drift can affect the decision making of managers and administrations of systems. An analysis of the event logs in BPMS allows the automatic detection of the concept drift. This paper presents an innovative method by introducing a modified distance function to identify the concept drift. Experimental results were performed on 72 datasets in the research history, which included 648 concept drifts in 12 different types. It shows that the proposed method detects 98.18% of the drifts, while the proposed method is much faster than other state of the art methods.

Keywords: Business process management systems**,** Process mining, Concept drift, Process drift detection