A Conceptual Framework for Preprocessing and Improving Quality of Event Log in Process Mining

Ahmad Salehi*
PhD Candidate in Information Technology Engineering; Tarbiat Modares University; Tehran, Iran Email: ahmad.salehi@modares.ac.ir

Mohammad Aghdasi
PhD in Industrial Engineering; Professor; Tarbiat Modares University; Tehran, Iran Email: aghdasim@modares.ac.ir

Toktam Khatibi
PhD in Industrial Engineering; Associate Professor; Tarbiat Modares University; Tehran, Iran Email: toktam.khatibi@modares.ac.ir

Majid Sheikhmohammady
PhD in Systems Design Engineering; Associate Professor; Tarbiat Modares University; Tehran, Iran Email: msheikhm@modares.ac.ir

Iranian Journal of Information Processing and Management

Iranian Research Institute for Information Science and Technology (IranDoc)
ISSN 2251-8223
eISSN 2251-8231
Indexed by SCOPUS, ISC, & LISTA
Vol. 38 | No. 3 | pp. 945-980
Spring 2023
https://doi.org/jipm.38.3

Received: 04, Jan. 2022 **Accepted: 22, Jun. 2022**

Abstract: In today's challenging world, organizational growth is not possible without the efficient use of data. Process mining uses machine learning methods and business process management concepts to extract hidden knowledge about business processes from data stored in information systems. Process Discovery is the first step in process mining. The main goal of process discovery is to transform the event log into a process model. However, using process discovery methods will not be possible without appropriate data because any analysis based on low-quality data will lead to poor insights and bad decisions that will negatively affect the performance of the organization or business. This paper aims to provide a new conceptual framework for preprocessing data input into process discovery methods to improve the quality of the extracted model. The proposed conceptual framework has been developed using a qualitative research process based on grounded theory. For this purpose, 102 articles related to the domain of data quality in process mining were reviewed, and the most critical challenges of data quality in this field have been identified after filtering and integrating them from the literature, including "noisy/infrequent events", "outlier events", "anomalous events", "missing values", "incorrect time format", "ambiguous timestamps", "synonymous activities", and "size and complexity". Then, the basic steps for data preprocessing and cleaning tasks are defined,

* Corresponding Author

which include the activities of “repair”, “anomaly detection”, “filtering”, and “dimensional reduction. The final preprocessing framework then builds on data quality issues and identified activities. Four standardized datasets derived from real-world processes were used to assess the proposed framework’s performance. Firstly, these data are raw, and secondly, four standard process discovery algorithms are applied after preprocessing by the introduced framework. The results showed that the preprocessing of the input data leads to the improvement of the model quality criteria extracted from the process discovery algorithms. Furthermore, to evaluate the validity of the proposed framework, its performance was compared with three preprocessing methods: “sampling”, “statistical preprocessing”, and “prototype selection”, which the results indicate better efficiency of the proposed approach. The results of this study can be used as guidelines by data and business analysts to identify and resolve data quality problems in process mining projects.

Keywords: Information Systems, Business Process Management, Process Mining, Data Quality, Event Log Preprocessing

ارائهٔ یک چارچوب مفهومی برای پیش‌پردازش و بهبود کیفیت نگاره‌های رویداد در فرایندکاوی

احمد صالحی
دانشجوی دکتری مهندسی فناوری اطلاعات؛
دانشگاه تربیت مدرس؛ تهران، ایران؛
پدیدآور رابط ahmad.salehi@modares.ac.ir

محمد اقدسی
دکتری مهندسی صنایع؛ استاد؛
دانشگاه تربیت مدرس؛ تهران، ایران؛
aghdasim@modares.ac.ir

توکتم خطیبی
دکتری مهندسی صنایع؛ دانشیار؛
دانشگاه تربیت مدرس؛ تهران، ایران؛
toktam.khatibi@modares.ac.ir

مجید شیخ‌محمدی
دکتری مهندسی طراحی سیستم؛ دانشیار؛
دانشگاه تربیت مدرس؛ تهران، ایران؛
msheikhm@modares.ac.ir

پژوهش‌نامهٔ پردازش و مدیریت اطلاعات

دریافت: ۱۴۰۰/۱۰/۱۴ | پذیرش: ۱۴۰۱/۰۴/۰۱ | مقاله برای اصلاح به مدت ۵۸ روز نزد پدیدآوران بوده است.

چکیده: در دنیای پیچیدهٔ امروز حیات سازمان‌ها و کسب‌وکارها بدون شناخت و استفادهٔ کارآمد از داده‌ها امکان‌پذیر نخواهد بود. فرایندکاوی با ترکیب روش‌های یادگیری ماشین و مفاهیم مدیریت فرایندهای کسب‌وکار تلاش دارد دانش نهان مربوط به چگونگی اجرای فرایندها را از داده‌های ذخیره‌شده در سامانه‌های اطلاعاتی استخراج نماید. اولین گام در فرایندکاوی، فعالیت کشف فرایند است که امکان مدل‌سازی فرایندها بر مبنای داده‌های رویداد ورودی را فراهم می‌سازد. اما استفاده از این مزیت بدون وجود داده‌های مناسب و باکیفیت فراهم نخواهد شد، زیرا هر گونه تحلیل بر پایهٔ داده‌های با کیفیت پایین منجر به ایجاد بینش و تصمیمات نامناسبی می‌شود که بر عملکرد سازمان یا کسب‌وکار تأثیر منفی خواهد گذاشت. هدف این پژوهش ارائهٔ یک چارچوب مفهومی جدید برای پیش‌پردازش داده‌های ورودی به روش‌های کشف فرایند است تا کیفیت مدل فرایند نهایی بهبود یابد. چارچوب مفهومی پیشنهادی با

نشریهٔ علمی | رتبهٔ بین‌المللی
پژوهشگاه علوم و فناوری اطلاعات ایران (ایرانداک)
شاپا (چاپی) ۸۲۲۳-۲۲۵۱
شاپا (الکترونیکی) ۸۲۳۱-۲۲۵۱
نمایه در LISTA، ISC، و SCOPUS
jipm.irandoc.ac.ir
دورهٔ ۳۸ | شمارهٔ ۳ | ص‌ص ۹۴۵-۹۸۰
بهار ۱۴۰۲
https://doi.org/jipm.38.3

استفاده از یک روش پژوهش کیفی بر اساس نظریهٔ داده‌بنیاد پدید آمده است. بدین منظور، ۱۰۲ پژوهش مرتبط با حوزهٔ کیفیت داده در فرایندکاوی مورد بررسی قرار گرفته و مهم‌ترین چالش‌های کیفیت داده در این زمینه پس از پالایش و یکپارچه‌سازی آن‌ها از ادبیات شناسایی شده‌اند که شامل: «رویدادهای آشفته/ کم‌تکرار»، «رویدادهای پرت»، «رویدادهای ناهنجار»، «مقادیر گمشده»، «قالب زمانی نادرست»، «برچسب‌های زمانی مبهم»، «فعالیت‌های مترادف» و «اندازه و پیچیدگی» است. در ادامه، گام‌های اساسی برای پیش‌پردازش و پاک‌سازی مناسب داده‌ها تعیین شده‌اند که دربرگیرندهٔ فعالیت‌های «ترمیم»، «کشف ناهنجاری»، «پالایش» و «کاهش ابعاد» می‌شوند. سپس، چارچوب مفهومی نهایی بر پایهٔ مشکلات کیفیت داده و فعالیت‌های پاک‌سازی شناسایی‌شده ایجاد شده است. برای بررسی عملکرد چارچوب پیشنهادی از چهار مجموعه دادهٔ استاندارد برگرفته از فرایندهای واقعی استفاده شده است. این داده‌ها در مرحلهٔ اول به‌صورت خام و در مرحلهٔ دوم پس از انجام پیش‌پردازش توسط چارچوب معرفی‌شده به چهار الگوریتم متداول کشف فرایند اعمال شده‌اند. نتایج نشان داد که پیش‌پردازش داده‌های ورودی منجر به بهبود معیارهای کیفیت مدل استخراج‌شده از الگوریتم‌های کشف فرایند می‌شود. همچنین، برای سنجش اعتبار چارچوب پیشنهادی، عملکرد آن با سه روش پیش‌پردازش «نمونه‌برداری»، «پیش‌پردازش آماری» و «انتخاب نمونهٔ اولیه» مقایسه شده که برایندها بیانگر کارایی بهتر رویکرد پیشنهادی بوده است. نتایج پژوهش حاضر می‌تواند به‌عنوان یک رهیافت کاربردی توسط متخصصان و تحلیلگران داده و کسب‌وکار در پروژه‌های فرایندکاوی مورد استفاده قرار گیرد.

کلیدواژه‌ها: سامانه‌های اطلاعاتی، مدیریت فرایندهای کسب‌وکار، فرایندکاوی، کیفیت داده، پیش‌پردازش نگاره رویداد

۱. مقدمه و بیان مسئله

با رشد فناوری و اهمیت پذیرش سازوکارهای تحول دیجیتال در پیشرفت سازمان‌ها و کسب‌وکارها، نقش سامانه‌های اطلاعاتی[1] در پیاده‌سازی و بازبینی فرایندها و همچنین مدیریت جریان اطلاعات و داده‌ها از اهمیت بسیار زیادی برخوردار شده است. به همین دلیل، می‌توان بیان نمود که امروزه، فرایندهای عملیاتی و سامانه‌های اطلاعاتی با یکدیگر در هم تنیده شده‌اند (Van Der Aalst 2016; Li 2020). داده‌های ذخیره‌شده در پایگاه داده‌های سامانه‌های اطلاعاتی دیدگاه لازم برای تصمیم‌گیری را فراهم نموده و به دوران تصمیم‌گیری بر پایهٔ احساس و شهود پایان داده‌اند. در حقیقت، کسب‌وکارهای امروزی به موجودیت‌های مبتنی بر داده تبدیل شده‌اند (شامی زنجانی، نبیبی و ایران‌دوست ۱۳۹۹). از

1. information systems

طرفی، باید بیان کرد که فرایندهای کسب‌وکار[1] بخشی از دارایی‌های اصلی یک سازمان محسوب شده و منجر به آگاهی از شرح وظایف، نقش‌ها و مسئولیت‌ها می‌شوند. به بیان دیگر، فرایندها فعالیت‌هایی را که برای خلق ارزش، خدمات و محصولات لازم است، شرح می‌دهند (دوماس و همکاران ۲۰۱۳).

برای بهره‌برداری از داده‌های ذخیره‌شده در پایگاه‌های داده در زمان اجرای فرایندهای کسب‌وکار، دانشی نوین تحت عنوان «فرایندکاوی»[2] توسعه یافته که از پیوند روش‌های یادگیری ماشین و داده‌کاوی با مفاهیم مدیریت فرایندهای کسب‌وکار[3] ایجاد شده است (فن در آلست ۲۰۱۱). هدف فرایندکاوی کشف، پایش و بهبود فرایندها با استفاده از بررسی و کاوش داده‌های برآمده از پیاده‌سازی و اجرای فرایندهاست.

نقطهٔ شروع برای هر پروژه یا مطالعهٔ فرایندکاوی، فعالیت «کشف فرایند»[4] است (Sani 2020). روش‌های فرایندکاوی از منابع داده‌ای ویژه‌ای به نام «داده‌های رویداد»[5] بهره می‌گیرند که در ادبیات از آن‌ها با نام «نگارهٔ رویداد»[6] یا به‌اختصار، «لاگ»[7] یاد می‌شود[8]. هر نگارهٔ رویداد ساختاری استاندارد برای ذخیره‌سازی رویدادهای ثبت‌شده در زمان اجرای یک فرایند است. الگوریتم‌های کشف فرایند با بررسی رویدادهای درون نگارهٔ رویداد سعی در یافتن روابط بین آن‌ها نموده و این‌گونه فرایند کسب‌وکار نهفته در داده‌های ذخیره‌شده را استخراج و مدل‌سازی می‌کنند. مدل فرایند کسب‌وکار حاصل، یک الگوی ترسیمی از وظایف، رویدادها و تصمیماتی است که به‌هنگام اجرای یک فرایند اتفاق افتاده و می‌توانند به شکل‌های مختلف از جمله شبکه‌های «پتری»[9] یا مدل و نشانگر فرایند کسب‌وکار[9] نمایش داده شوند (Burattin 2015). شایان توجه است که استفاده از روش‌های فرایندکاوی بدون وجود نگاره‌های رویداد امکان‌پذیر نخواهد بود (Batyuk and Voityshyn 2018). شکل ۱، نمونه‌ای از یک نگارهٔ رویداد و مدل فرایند استخراج‌شده از آن توسط روش‌های کشف فرایند را به تصویر کشیده است. هر ردیف نگارهٔ رویداد متناظر است با یک رویداد انجام‌شده در زمان اجرای فرایند و ستون‌ها نشان‌دهندهٔ ویژگی‌های مربوط به آن رویداد است که شامل موارد زیر است:

---

1. business processes
2. process mining
3. business process management (BPM)
4. process discovery
5. event data
6. event log
7. log

۸. در بعضی از متون فارسی به‌جای عبارت «نگارهٔ رویداد»، از معادل «گزارش رویداد» استفاده شده است.

9. petri net
9. Business Process Model and Notation (BPMN)

۱. شناسهٔ رویداد[1]: برای تمایز بین رویدادهای ذخیره شده در یک نگارهٔ رویداد و همچنین، به‌منظور ارجاع به یک رویداد خاص استفاده می‌شود؛
۲. شناسهٔ مورد[2]: بیانگر این است که هر رویداد به کدام نمونه یا مورد که فرایند برای آن در جریان است، تعلق دارد؛
۳. فعالیت[3]: برای تعیین اینکه در یک رویداد خاص، کدام فعالیت انجام شده است؛
۴. برچسب زمانی[4]: نشان‌دهندهٔ زمان و تاریخ انجام یک فعالیت برای یک رویداد مشخص است.

شکل ۱. نمونه‌ای از یک نگارهٔ رویداد و مدل کشف‌شده بر اساس داده‌های ذخیره‌شدهٔ درون آن

هر نگارهٔ رویداد می‌تواند شامل ویژگی‌های بیشتری باشد، اما برای فرایندکاوی وجود سه ویژگی فوق الزامی است (Van Der Aalst 2016). همچنین، به مقادیر ذخیره‌شده برای هر ویژگی، برچسب[5] گفته می‌شود. همان‌طور که در نگارهٔ رویداد شکل ۱، مشاهده می‌شود، این فرایند دارای پنج فعالیت اصلی است که شامل وظایف "a,b,c,d,e" هستند. لازم به ذکر است که به توالی متناهی از فعالیت‌هایی که برای هر مورد انجام شده‌اند، یک مسیر[6] یا دنباله[7] می‌گویند. به‌عنوان نمونه، برای مورد ۱، مسیر فرایند یا دنباله به‌صورت {a,b,c,d,e} تعریف می‌شود.

روش‌های کشف فرایند همواره فرض می‌کنند که داده‌های ورودی کامل و بدون خطا هستند، اما ممکن است مدل استخراج‌شده به‌دلیل وجود داده‌های آشفته[8]، نامناسب و نادرست نتواند فرایند را به‌صورت صحیح مدل‌سازی کند. به‌عبارت بهتر کیفیت پایین داده‌های ورودی منجر به ایجاد بینش و دانش اشتباه خواهد شد (Pyle 1999). در بیشتر موارد، روش‌های فرایندکاوی بر روی مجموعه داده‌های مصنوعی به‌خوبی عمل می‌کنند، اما به‌هنگام استفاده از داده‌های واقعی منجر به ایجاد مدل‌های کسب‌وکار پیچیده‌ای

1. event_id
2. case_id
3. activity
4. timestamp
5. label
6. trace
7. sequence
8. noise

خواهند شد که درک آن‌ها به‌آسانی امکان‌پذیر نیست (Sani 2020). بدین‌سان، هر نگارهٔ رویداد همواره نیازمند پیش‌پردازش[1] است، تا رویدادها به‌صورت مناسب پاک‌سازی شده و قابلیت بهره‌برداری از آن‌ها افزایش یابد. از طرفی، چون ممکن است این عمل در دنیای واقعی به‌صورت دستی انجام شود، فعالیت پیش‌پردازش به تجربه و دانش تحلیلگر بسیار وابسته خواهد شد (Andrews et al. 2020). «امام جمعه» و همکاران با بررسی و تحلیل بیش از ۱۵۰ نمونه از پروژه‌های فرایندکاوی به این نتیجه رسیدند که فعالیت‌های مربوط به بررسی کیفیت داده‌نگاره‌های رویداد و پیش‌پردازش آن‌ها به‌صورت بسیار سطحی انجام پذیرفته است (Emamjome et al. 2020)، که این مورد افزون بر افزایش هزینه و زمان، اعتبار نتایج خروجی را تحت تأثیر قرار خواهد داد.

بنابراین، نیاز به فرایند پیش‌پردازش مناسب که هدف آن تبدیل داده‌های ورودی ثبت‌شده از دنیای واقعی به داده‌های استاندارد و باکیفیت است که نتایج خروجی حاصل از روش‌های کشف فرایند را تحت تأثیر قرار ندهد. این مهم یک فعالیت کلیدی برای دستیابی به نتایج قابل اعتماد و کاهش هزینه‌ها خواهد بود. انگیزهٔ اصلی پژوهش حاضر معرفی یک چارچوب مفهومی[2] جامع و مشخص برای پالایش داده‌های ورودی به روش‌های کشف فرایند در زمینهٔ فرایندکاوی است؛ به‌گونه‌ای که مهم‌ترین چالش‌های کیفیت داده را شناسایی نموده و تأثیر آن‌ها را بر روی نتایج خروجی به حداقل برساند. بر پایهٔ هدف پژوهش، این مطالعه به‌دنبال پاسخگویی به سه سؤال زیر است:

۱. آیا بدون دانش قبلی و مدل فرایند کسب‌وکار مرجع (مستندات سازمان یا کسب‌وکار)، می‌توان عملیات پیش‌پردازش بر روی داده‌های رویداد را انجام داد؟
۲. آیا با شناسایی مهم‌ترین مشکلات کیفیت داده و استفاده از پیش‌پردازش مناسب، امکان بهبود نتایج خروجی الگوریتم‌های کشف فرایند فراهم می‌شود؟
۳. آیا می‌توان فعالیت‌های مهم در بهبود کیفیت نگارهٔ رویداد را با یکدیگر یکپارچه نمود؟

ساختار این مقاله بدین شرح است: بخش ۲، بیان‌کنندهٔ پیشینهٔ پژوهش‌های انجام‌شده در حوزهٔ کیفیت داده و پیش‌پردازش نگاره‌های رویداد است. بخش ۳، روش انجام این

1. preprocessing
2. conceptual framework

پژوهش را شرح خواهد داد. بخش ۴، به تشریح یافته‌های این پژوهش می‌پردازد و در بخش ۵، خلاصه‌ای از نتایج و پیشنهاداتی برای مطالعات آتی مطرح می‌شوند.

۲. پیشینهٔ پژوهش

در ادبیات فرایندکاوی عبارت «اگر داده‌های نامعتبر و نادرست به سامانه وارد شود، خروجی آن نیز نامعتبر خواهد بود»[1]، فراوان مشاهده می‌شود (Verhulst 2016; Suriadi et al. 2017; Andrews et al. 2018; Sani 2020). این امر بیانگر اهمیت کیفیت داده‌های ورودی (نگارهٔ رویداد) به روش‌های فرایندکاوی است. به همین منظور، پژوهش‌هایی برای ایجاد مفاهیم نظری کیفیت داده در فرایندکاوی با هدف شناخت بهتر چالش‌های این زمینه انجام شده است. برای نخستین‌بار، در «مانیفست فرایندکاوی»[2] به سطوح بلوغ کیفیت نگاره‌های رویداد اشاره می‌شود که آن‌ها را در ۵ سطح از عالی تا ضعیف دسته‌بندی نموده است (Van Der Aalst et al. 2011). در جایی دیگر، «بوس، منز و فن در آلست» تشریح دقیق‌تری از مشکلات داده‌های رویداد بر اساس تجارب گذشتهٔ خود در پروژه‌های فرایندکاوی را بیان می‌کنند. آن‌ها مشکلات کیفیت نگاره‌های رویداد را در ۱۰ گروه متفاوت دسته‌بندی نموده و برای هر دسته، ویژگی‌ها و چگونگی تأثیر آن‌ها بر روی نتایج خروجی را به‌صورت کامل شرح می‌دهند (Bose, Mans, and van der Aalst 2013). در مقاله‌ای دیگر، «سوریادی» و همکاران، الگوهای رایج ایجاد خطا در نگاره‌های رویداد را برشمرده و موفق شدند ۱۱ الگوی نقص داده را یافته و با ارائهٔ مثال‌هایی نقش این الگوها را در فرایندکاوی روشن سازند (Suriadi et al. 2017). مطالعات فوق با وجود روشنگری و دسته‌بندی مشکلات کیفیت داده در فرایندکاوی، راهکاری برای چگونگی پیش‌پردازش داده‌ها ارائه نمی‌دهند. همچنین، بعضی از مطالعات در رویکردی متفاوت تلاش نموده‌اند روش‌های کشف فرایند را نسبت به وجود داده‌های پرت و آشفته مقاوم نمایند تا این الگوریتم‌ها بتوانند یک مدل فرایند کسب‌وکار صحیح را از داده‌های ورودی با کیفیت پایین استخراج نمایند. نمونه‌ای از این تلاش‌ها در مطالعات «لیمانز، فهلند و فن در آلست» و همچنین «آگوستو» و همکاران به مرحلهٔ اجرا درآمده است (Leemans, Fahland and van der Aalst 2013; Augusto et al. 2019). اما این روش‌ها در دستهٔ الگوریتم‌های کشف فرایند

1. garbage in, garbage out
2. process mining manifesto

قرار گرفته و از تنظیمات داخلی برای پالایش بعضی از رویدادها بهره می‌برند و نمی‌توانند به‌عنوان یک راهکار پیش‌پردازش در نظر گرفته شوند. بنابراین، در سال‌های اخیر، توسعه و استفاده از رویکردهای پیش‌پردازش داده‌ها در فرایندکاوی مورد توجه قرار گرفته است.

مرور پژوهش‌های مربوط به پیش‌پردازش داده‌ها در فرایندکاوی نشان داد که مطالعات انجام‌پذیرفته در این زمینه را می‌توان بر اساس نوع هدف به چهار رویکرد متفاوت دسته‌بندی نمود. دستهٔ اول، رویکرد خود را بر مبنای شناسایی و حذف مسیرهای فرایندی که میزان تکرار آن‌ها در داده‌های ورودی کمتر از میزان یک آستانهٔ مشخص بود، قرار داده‌اند. به این ترتیب، با حذف این مسیرها نتایج خروجی بهبود می‌یابند. اولین تلاش در این زمینه در مطالعه‌ای توسط «لی» و همکاران انجام پذیرفت. در این پژوهش نویسندگان با شناسایی مسیرهای فرایند کم‌تکرار یا آشفته و حذف آن‌ها از لاگ موفق شدند کیفیت مدل فرایند استخراجی را افزایش دهند (Ly et al. 2012). در پژوهشی دیگر، «تَکس، سیدوروا و فن در آلست» با تعریف نظری فعالیت‌های آشفته و بی‌نظم به همراه چگونگی شناسایی و حذف آن‌ها از مجموعه داده‌های ورودی تلاش نمودند کیفیت مدل کشف‌شده را بهبود دهند (Tax, Sidorova and van der Aalst 2019). در مقالهٔ «لو» و همکاران، رویکردی جدید برای حذف رویدادهای آشفته و کم‌تکرار پیشنهاد شده است. نویسندگان با ارائهٔ یک راه‌حل جدید به دنبال ایجاد تمایز بین رفتارهای کم‌تکرار ناشی از شرایط خاص یک فرایند، با رویدادهای غیرمفید یا آشفته بوده‌اند تا مدل نهایی بدین‌گونه بازنمایی درستی از اجرای یک فرایند کسب‌وکار را ارائه دهد (Lu et al. 2021).

دستهٔ دیگر پژوهش‌ها به بررسی و شناسایی رویدادهای تکراری پرداخته‌اند. برای مثال، در تحقیق «لو» و همکاران، یک روش پیش‌پردازش به‌منظور شناسایی و پالایش رویدادهای تکراری ارائه شده است. در این تحقیق با یکپارچه‌سازی رویدادهای تکراری، کیفیت نتایج خروجی بهبود یافته‌اند (Lu et al. 2016).

دستهٔ سوم پژوهش‌های پیش‌پردازش رهیافت خود را بر مبنای شناسایی و حذف دنباله‌های ناقص یا رویدادهای پرت استوار ساخته‌اند. در مطالعات «آیو، فولورونسو و ایبهارالو» و همچنین «ثانی، فن زلست و فن در آلست» با استفاده از رویکردهایی که بر مبنای احتمالات شرطی بنیان شده‌اند، دنباله‌های ناقص یا رویدادهای پرت شناسایی و از مجموعه داده‌های ورودی حذف می‌شوند (Ayo, Folorunso and Ibharalu 2017; Sani, van Zelst and van der Aalst 2017). حالت تکامل‌یافتهٔ این رویکرد در پژوهش دیگری معرفی

شده است. در این تحقیق روشی سه-مرحله‌ای توسعه داده شده که در گام اول، قوانین و الگوها از دنباله‌های درون یک لاگ استخراج می‌شوند. در گام دوم، نقاط پرت برای هر دنباله شناسایی شده و سپس، در گام سوم، دنباله‌های دارای رفتارهای پرت از لاگ حذف می‌شوند (Sani, van Zelst, and van der Aalst 2018a).

در دستهٔ چهارم مطالعات بررسی‌شده به کاهش ابعاد و پیچیدگی داده‌های ورودی به‌عنوان یک فعالیت پیش‌پردازش در فرایندکاوی پرداخته شده است. در پژوهش «بویر» و دیگران، یک چارچوب برای بهبود کشف فرایند از طریق یک روش «پیش‌پردازش آماری»[1] به‌منظور کاهش اندازه و پیچیدگی نگارهٔ رویداد ارائه شده است. هدف این مطالعه کاهش حجم حافظه و زمان اجرای الگوریتم‌های کشف فرایند است (Bauer et al. 2018). «ثانی، بولتنهاگن و فن در آلست» در مطالعه‌ای دیگر، روش پیش‌پردازشی به‌نام «انتخاب نمونهٔ اولیه»[2] معرفی نمودند که شامل چهار مرحلهٔ خوشه‌بندی نمونه‌های اولیه ورودی، کشف مدل بر اساس نمونه‌های داخل خوشه‌ها، ارزیابی نتایج خروجی، و در انتها، تکرار سه مرحلهٔ قبلی تا کسب بهترین نتیجه است (Sani, Boltenhagen and van der Aalst 2019). «ثانی، فن زلست و فن در آلست» در ادامهٔ پژوهش‌های خود، روش «نمونه‌برداری»[3] را توسعه داده‌اند که هدف آن انتخاب زیرمجموعه‌ای از داده‌های رویداد است که بیشترین همپوشانی را با فرایند کسب‌وکار مرتبط با آن‌ها را داشته باشد. این عمل افزون بر کاهش اندازهٔ نگارهٔ رویداد منجر به افزایش کیفیت مدل خروجی از روش‌های کشف فرایند خواهد شد (Sani,van Zelst and van der Aalst 2019). در جدول ۱، پژوهش‌های مرتبط با پیش‌پردازش داده‌های ورودی در فرایندکاوی بر پایهٔ نوع رویکرد و هدف آن‌ها به‌صورت خلاصه بیان شده است.

1. statistical pre-processing
2. prototype selection
3. sampling

جدول ۱. پژوهش‌های مربوط به زمینهٔ پیش‌پردازش نگاره‌های رویداد در فرایندکاوی

نوع رویکرد	هدف	منبع
شناسایی و حذف مسیرهای کم‌تکرار یا آشفته	ارائهٔ چارچوبی به نام «پالایش معنایی لاگ»[1] به‌عنوان یک مرحلهٔ پیش‌پردازش برای پاک‌سازی نگارهٔ رویداد بر اساس محدودیت‌های مربوط به زمینهٔ کسب‌وکار. این چارچوب اقدام به حذف مسیرهای کم‌تکرار یا آشفته از نگارهٔ رویداد ورودی می‌نماید. هدف این چارچوب افزایش دقت مدل فرایند استخراج‌شده از لاگ است.	Ly et al. (2012)
	این پژوهش به موضوع چگونگی کشف فرایند بر مبنای نگاره‌های رویداد دارای مشکل کیفیت داده می‌پردازد که در آن چند روش نوین برای پیش‌پردازش داده‌های ورودی معرفی می‌شوند. هدف روش معرفی‌شده حذف فعالیت‌های آشفته و بی‌نظم از نگارهٔ رویداد و در نتیجه، بهبود مدل فرایند کسب‌وکار استخراج‌شده است.	Tax, Sidorova, and van der Aalst (2019)
	ارائهٔ یک روش جدید برای تمایز بین رویدادهای آشفته از رویدادهای کم‌تکرار مهم و مؤثر در فرایند کسب‌وکار که بر مبنای یک مفهوم احتمالی به نام مسیرهای حداکثر احتمال[2] عمل می‌نماید. هدف این روش بررسی حالت‌های مختلف فعالیت‌ها در یک فرایند و همچنین، مشخص کردن حالت‌های گذار از یک فعالیت به فعالیت دیگر (تمایز بین فعالیت‌های منظم و نامنظم) است، تا رفتارهای موجود در یک دنباله را به دو قسمت سودمند و غیرمفید تقسیم کند.	Lu et al. (2021)
شناسایی و حذف رویدادهای تکراری	حل مشکل برچسب‌های نامناسب (در اینجا تکراری) با استفاده از یک روش پیش‌پردازش جدید که برچسب فعالیت‌های تکراری را با یکدیگر ادغام می‌کند.	Lu et al. (2016)
شناسایی و حذف دنباله‌های ناقص یا رویدادهای پرت	بررسی وابستگی بین فعالیت‌ها و ارائهٔ یک روش احتمالی برای پیش‌بینی فعالیت‌های گمشده در دنباله‌های ناقص با استفاده از توابع امتیازدهی بیزی[3]	Ayo, Folorunso, and Ibharalu (2017)
	توسعهٔ یک روش پالایش بر پایهٔ احتمالات شرطی[4] برای تعیین احتمال وقوع یک فعالیت بر اساس رفتار فعالیت‌های نزدیک به آن. اگر احتمال یک فعالیت کمتر از مقدار آستانه باشد، آن فعالیت به‌عنوان رفتار پرت در نظر گرفته می‌شود.	Sani, van Zelst, and van der Aalst (2017)
	معرفی روشی سه‌مرحله‌ای برای استخراج قوانین و الگوهای دنباله‌های یک لاگ، تا با شناسایی نقاط پرت برای هر دنباله و سپس حذف دنباله‌ها یا رفتارهای پرت از نگارهٔ رویداد، کیفیت مدل کشف‌شدهٔ نهایی افزایش یابد.	Sani, van Zelst, and van der Aalst (2018a)

1. semantic log purging
2. maximum probability path
3. bayesian scoring functions
4. conditional probabilities

نوع رویکرد	هدف	منبع
کاهش ابعاد و پیچیدگی	معرفی یک روش پیش‌پردازش آماری که تلاش دارد با انتخاب نمونه‌ای مناسب (زیرمجموعه) از داده‌های موجود در نگارهٔ رویداد که بیشترین همپوشانی را با فرایند کسب‌وکار مربوط به آن را داشته باشد، پیدا نموده و به این صورت به کاهش اندازه و پیچیدگی داده‌های ورودی منتهی شود و معیارهای کیفیت مدل نهایی افزایش یابد.	Bauer et al. (2018)
	ارائهٔ یک روش پیش‌پردازش چهار–مرحله‌ای شامل خوشه‌بندی دنباله‌های موجود در نگارهٔ رویداد برای ساخت نمونه‌های اولیهٔ ورودی، سپس کشف مدل بر اساس نمونه‌های ساخته‌شده، و آنگاه ارزیابی کیفیت نتایج خروجی و در انتها تکرار سه روند قبلی تا کسب بهترین نتیجه بر اساس معیارهای ارزیابی روش‌های کشف فرایند.	Boltenhagen, and van der Aalst (2019)
	توسعهٔ روشی که داده‌های رویداد ورودی را دریافت و یک نگارهٔ رویداد ساده‌شده را به‌عنوان خروجی تحویل می‌دهد. در این روش با استفاده از چهار استراتژی متفاوت شامل طول، تعداد تکرار، شباهت و ساختار، مسیرهای فرایند از داده‌های رویداد ورودی استخراج می‌شوند.	Fani Sani, Zelst, and van der Aalst (2019)

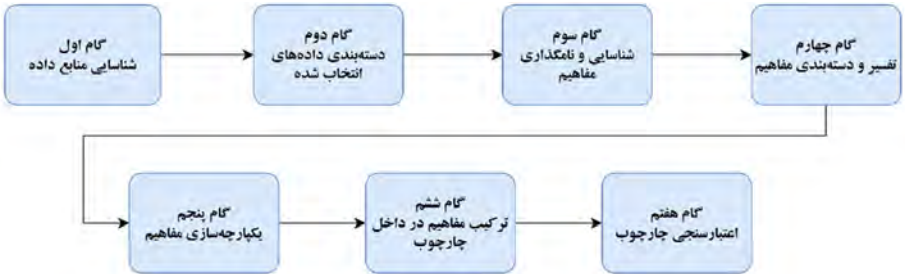
با وجود اینکه همهٔ روش‌های اشاره‌شده در جدول ۱، به موفقیت‌های مهمی در بهبود کیفیت مدل‌های کسب‌وکار استخراجی از داده‌های رویداد رسیده‌اند، اما با توجه به اینکه هر کدام از این روش‌ها بر اساس یک هدف مشخص توسعه یافته‌اند، موفق شده‌اند تعداد محدودی از چالش‌های کیفیت داده را حل نمایند. در رویکردهای فوق، حذف دنباله‌هایی با فعالیت‌های آشفته، تکراری یا دارای فعالیت پرت به‌صورت یک رویکرد پاک‌سازی جامع که شامل همهٔ موارد فوق باشد، مشاهده نشده است؛ یا همراستا با کاهش اندازه و پیچیدگی نگارهٔ رویداد ورودی فقط به حذف مسیرهایی که کم‌تکرار یا پرت هستند، پرداخته شده و ضرورت بازسازی و ترمیم مسیرهای دارای مشکل در نظر گرفته نشده است. به همین دلیل، معرفی رویکردی جامع و یکپارچه که توانایی شناسایی و حل مهم‌ترین چالش‌های کیفیت داده را داشته باشد، به‌عنوان یک شکاف پژوهش قابل بیان است. همچنین، معرفی رویکردی که بتواند به‌عنوان یک شیوه‌نامهٔ اجرایی توسط متخصصان فرایند استفاده شود، به‌عنوان دیگر شکاف پژوهش باید مورد توجه قرار گیرد.

۳. روش پژوهش و ایجاد چارچوب مفهومی

پژوهش حاضر از نظر هدف، کاربردی است و انگیزهٔ اصلی آن معرفی و توسعهٔ یک چارچوب مفهومی برای حل چالش‌های کیفیت داده در روش‌های کشف فرایند در زمینهٔ فرایندکاوی است. چارچوب مفهومی ابزاری است که امکان واکاوی و سازماندهی اطلاعات

مربوط به یک زمینهٔ دانش را در یک ساختار روشن فراهم نموده و انجام تحقیقات آتی در آن زمینه را آسان‌تر می‌نماید (Succar 2009; Zamora-Polo et al. 2019). بر پایهٔ تعریفی دیگر، هدف از یک چارچوب مفهومی توسعه و سازماندهی دانش موجود در مورد مفاهیم، مسئله‌ها یا مشکلات حوزهٔ مورد مطالعه است (Rocco and Plakhotnik 2009). بنابراین، استفاده از مزایای چارچوب‌های مفهومی برای شناخت مهم‌ترین چالش‌های کیفیت داده، چگونگی ارتباط و ارزیابی آن‌ها در داده‌های ورودی به همراه معرفی رهیافت‌هایی برای حل مشکلات فوق در زمینهٔ کشف فرایند می‌تواند بسیار سودمند باشد. همچنین، این کوشش می‌تواند دیدگاه‌های نوینی برای پژوهش‌های آینده در این زمینه ترسیم نماید.

در این مطالعه برای ایجاد چارچوب مفهومی همراستا با هدف پژوهش، از روش معرفی شده توسط «جبارین» استفاده شده است (Jabareen 2009). «جبارین» یک روش کیفی به‌منظور ایجاد چارچوب‌های مفهومی برای پدیده‌هایی که به مجموعه‌های دانش چندرشته‌ای مرتبط هستند، پیشنهاد نموده است. در این رویکرد، «مفهوم» به‌عنوان یک مؤلفه که شامل اجزایی خاص است، تعریف شده و از «چارچوب مفهومی» به‌عنوان یک شبکه از مفاهیم مرتبط به‌هم یاد می‌شود. سپس، با معرفی روش «واکاوی چارچوب مفهومی» که نوعی فرایند نظریه‌پردازی برای ساخت چارچوب مفهومی بر پایهٔ روش نظریهٔ داده‌بنیاد است، گام‌های عملی ایجاد چارچوب مفهومی شرح داده شده است. بدین‌سان برای ساخت یک چارچوب مفهومی، ابتدا باید پیشینهٔ موضوع پژوهش مورد بررسی قرار گرفته و مفاهیم باارزش از ادبیات استخراج شوند. سپس، چارچوب مفهومی بر اساس مفاهیم شناسایی شده و روابط بین آن‌ها ایجاد شود. شکل ۲، نمایی از مراحل ایجاد چارچوب مفهومی بر مبنای روش «جبارین» با هدف پیش‌پردازش نگاره‌های رویداد را نشان می‌دهد که در ادامه، فعالیت‌های مربوط به هر گام شرح داده خواهد شد.

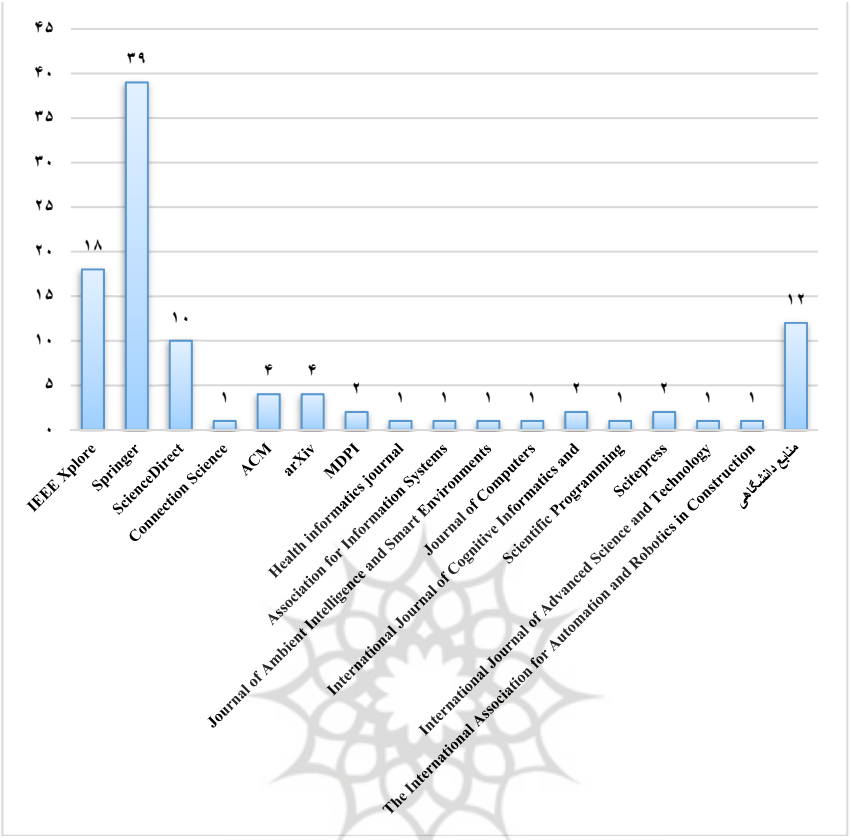


شکل ۲. مراحل اجرای ایجاد چارچوب مفهومی بر پایهٔ روش «جبارین» (Jabareen 2009)

۳-۱. شناسایی منابع داده

اولین مرحله برای ساخت یک چارچوب مفهومی، شناخت موارد بیان‌شده در مورد پدیدهٔ مورد نظر در ادبیات است. برای دستیابی به این هدف، پایگاه‌های علمی Web Of Science، Scopus، ScienceDirect، Springer، IEEE Xplore، Google Scholar به‌عنوان پایگاه‌های علمی معتبر مورد جست‌وجو قرار گرفتند. برای یافتن مقالات همراستا با هدف پژوهش از کلمات کلیدی مانند «کیفیت داده و فرایندکاوی»، «کیفیت نگارهٔ رویداد»، «پیش‌پردازش در فرایندکاوی»، «ارزیابی کیفیت داده و فرایندکاوی»، «کشف ناهنجاری[1] در نگارهٔ رویداد» و «بازسازی و ترمیم[2] نگارهٔ رویداد» استفاده شده است. به این ترتیب، تعداد ۱۰۲ مقالهٔ منتشرشده در همایش‌ها، نشریات، پایان‌نامه‌ها و گزارشات دانشگاهی شناسایی شدند که بیشترین پژوهش‌ها در سال‌های ۲۰۱۹ و ۲۰۲۰ میلادی انتشار یافته‌اند. مطابق انتظار، بیشتر مقالات به‌ترتیب در سه پایگاه علمی Springer، IEEE Xplore و Science Direct منتشر شده‌اند. مقالات چاپ‌شده در سه پایگاه فوق ۶۶ درصد از کل مطالعات انجام‌شده را دربرگرفته‌اند. در شکل ۳، تعداد پژوهش‌های منتخب انتشاریافته در پایگاه‌های علمی و نشریات مشخص شده است. شایان توجه است که در جست‌وجوهای انجام‌شده پژوهش‌های نگارش‌شده به زبان فارسی یافت نشد. همچنین، نتایج حاصل از این مرور نظام‌مند در پژوهش دیگری منتشر شده است.

1. anomaly detection
2. repair



شکل ۳. تعداد پژوهش‌های انتشاریافته در پایگاه‌های علمی و نشریات

۳-۲. دسته‌بندی داده‌های انتخاب‌شده

در این مرحله منابع انتخاب‌شده باید مطالعه و دسته‌بندی شوند. این کار با ترکیب قضاوت‌هایی کیفی در مورد مکان انتشار، اصالت نویسنده (نویسندگان)، جدید بودن و همچنین، بررسی تعداد استنادها انجام شده است. گفتنی است که از مجموع ۱۰۲ پژوهش بررسی‌شده، ۵۲ پژوهش در همایش‌ها ارائه شده و ۳۷ پژوهش در مجلات انتشار یافته‌اند که به‌ترتیب، ۵۱ درصد و ۳۶ درصد از کل پژوهش‌های مورد بررسی را شامل می‌شوند. همچنین، ۵ پایان‌نامه و ۸ گزارش دانشگاهی در این زمینه انجام گرفته است که سهم آن‌ها برابر با ۱۳ درصد مطالعات کاوش‌شده است. با کاوش مطالعات انتخاب‌شده، بر اساس نوع آن‌ها چهار رویکرد اصلی در زمینهٔ مشکلات کیفیت داده در فرایندکاوی شناسایی گردید

کـه شـامل «چارچوب‌هـای ارزیابـی کیفیـت داده»، «پیش‌پـردازش»، «ترمیـم» و «کشـف ناهنجـاری» هسـتند. در جـدول ۲، تعـداد مطالعـات مربـوط بـه هـر کـدام از رویکردهـای اشاره‌شـده بیـان شـده اسـت.

جدول ۲. تعداد منابع مطالعه‌شده به‌تفکیک نوع آن‌ها

نوع دسته‌بندی	تعداد مقالات همایشی	تعداد مقالات منتشرشده در مجلات	تعداد پایان‌نامه‌ها	مجموع
چارچوب‌های ارزیابی کیفیت داده‌ها	۱۱	۷	۲	۲۰
پیش‌پردازش	۲۱	۱۵	۲	۳۸
ترمیم	۱۳	۹	۱	۲۳
کشف ناهنجاری	۱۴	۷	–	۲۱
مجموع	۶۰	۳۸	۵	۱۰۲

۳-۳. شناسایی و نامگذاری مفاهیم

هـدف ایـن گام مطالعـهٔ منابـع انتخاب‌شـده اسـت تـا مفاهیـم مـورد نیـاز اسـتخراج و نام‌گـذاری شـوند. در طـی ایـن فعالیـت، ۱۰۲ مقالـهٔ مرتبـط بـا هـدف پژوهـش مطالعـه و بازخوانـی شـده‌اند. بدین‌ترتیـب، مهم‌تریـن مفاهیـم مربـوط بـه مشـکلات رایـج کیفیـت داده کـه بـر روی کشـف فراینـد تأثیرگـذار هسـتند، بـه همـراه تعاریـف هـر کـدام از آن‌هـا اسـتخراج شـده و به‌شـرح زیـر هسـتند:

- ◇ مسـیرهای آشـفته/ کم‌تکـرار: بـه شـرایطی اشـاره دارد کـه یـک یـا تعـدادی از مسـیرهای فراینـد، در نـگارهٔ رویـداد کمتـر از دیگـر مسـیرها اتفـاق افتاده‌انـد. وجـود این‌گونـه مسـیرها منجـر بـه ایجـاد ناهنجـاری در یـک لاگ خواهـد شـد (Ghionna et al. 2008)؛
- ◇ رویدادهـای بـدون برچسـب[1]: زمانـی کـه یـک رویـداد در نـگارهٔ رویـداد ثبـت شـده، امـا شناسـهٔ مـورد آن مشـخص نباشـد، بیـان می‌شـود کـه نـگارهٔ رویـداد دارای رویـداد بـدون برچسـب اسـت (Ferreira and Gillblad 2009; Helal and Awad 2020)، یعنـی نمی‌تـوان یـک رویـداد را بـه یـک نمونـهٔ مـورد خـاص نسـبت داد؛
- ◇ رویدادهـای آشـفته/ کم‌تکـرار: رویدادهـای کم‌تکـرار یـا نامتـداول رویدادهایـی هسـتند

1. unlabeled events

که به‌ندرت در جریان فرایند رخ می‌دهند (Bose, Mans and van der Aalst 2013). همچنین، در این بخش با نوع دیگری از رفتارها هم روبه‌رو هستیم که تحت عنوان آشفته از آن‌ها یاد می‌شود. رفتارهای بی‌نظم گونه‌ای از آشفتگی هستند که هیچ وابستگی به اصل فرایند کسب‌وکار ندارند و می‌توانند در هر بخش از فرایند رخ دهند (Yi and Peng 2019)؛

- رویدادهای گمشده[1]: بیانگر شرایطی است که یک رویداد در واقعیت در هنگام اجرای یک فرایند کسب‌وکار رخ دهد، اما در نگارهٔ رویداد ثبت نشود (Van Der Aalst 2016)؛
- ویژگی‌های گمشده: نشان‌دهندهٔ وضعیتی است که ممکن است شناسهٔ مورد، فعالیت یا برچسب زمانی یک رویداد مشخص‌نشده باشد (Sim, Bae and Choi 2019)؛
- برچسب‌های زمانی مبهم/ نادرست[2]: سطح انتزاع مقدار ثبت‌شده بازهٔ زمانی و تاریخی انجام یک رویداد را به‌درستی مشخص نمی‌سازد یا مقدار ثبت‌شده نادرست است (Van Der Aalst 2016)؛
- قالب[3] زمانی نادرست: این چالش زمانی ایجاد می‌شود که ساختار داده‌ای برای ذخیره‌سازی برچسب زمانی روش‌های کشف فرایند مناسب نباشد (Conforti, La Rosa, and Ter Hofstede 2018)؛
- فعالیت‌های بی‌نظم[4]: هر فعالیتی که به‌صورت ناگهانی (بدون برنامه‌ریزی قبلی) در هر نقطه از اجرای فرایند رخ دهد (Tax, Sidorova and van der Aalst 2019)؛
- فعالیت‌های تکراری[5]: به حالتی اشاره دارد که یک فعالیت ممکن است دو یا بیش از دو بار در یک فرایند تکرار شود که در این حالت الگوریتم‌های کشف فرایند ممکن است یک حلقه پیرامون آن فعالیت ایجاد نمایند (Lu, Fahland and van der Aalst 2016)، در صورتی که بنا به‌دلایلی ممکن است یک فعالیت مشخص چندین مرتبه در یک فرایند تکرار شود؛
- فعالیت‌های پراکنده: زمانی رخ می‌دهد که فعالیت‌های ذخیره‌شده برای یک فرایند از یکدیگر بسیار متفاوت هستند (Suriadi et al. 2017; Sadeghianasl et al. 2019)؛
- فعالیت‌های مترادف: به حالتی اشاره دارد که برچسب فعالیت‌ها از نظر معنایی با

1. missing events
2. imprecise/ incorrect
3. format
4. chaotic activities
5. duplicate

یکدیگر مترادف هستند (Sadeghianasl et al. 2019; Chen et al. 2021) و می‌توان آن‌ها را یک فعالیت واحد در نظر گرفت؛

◇ اندازه و پیچیدگی[1]: این مورد به بزرگ بودن اندازهٔ داده‌های رویداد ورودی به همراه حجم عظیمی از جزئیات غیرضروری اشاره دارد (Bose, Mans and van der Aalst 2013).

۳-۴. تفسیر و دسته‌بندی مفاهیم

هدف این مرحله شناخت و واکاوی عمیق مفاهیم و دسته‌بندی آن‌هاست. برای دستیابی به این هدف چالش‌های کیفیت داده در فرایندکاوی با توجه به سطحی که ممکن است در آن رخ دهند، به چهار سطح «نگارهٔ رویداد»، «مسیر فرایند»، «رویداد» و «ویژگی رویداد» تقسیم شده‌اند. همچنین، چالش‌های کیفیت داده مربوط به هر دسته که بر روی کیفیت آن سطح اثرگذار هستند، در جدول ۳، مشخص شده‌اند. این گام باعث می‌شود که چالش‌های کیفیت داده دسته‌بندی شده و ویژگی‌های مشترک آن‌ها شناسایی شود.

جدول ۳. مشکلات کیفیت داده به‌تفکیک هر سطح

سطح	مشکلات کیفیت داده
نگارهٔ رویداد	اندازه و پیچیدگی
مسیر فرایند	مسیرهای کم‌تکرار/ آشفته
رویداد	رویدادهای گمشده، رویدادهای آشفته
ویژگی رویداد	ویژگی‌های گمشده، رویدادهای بدون برچسب، فعالیت‌های تکراری، فعالیت‌های بی‌نظم، فعالیت‌های ناخالص، فعالیت‌های مترادف، برچسب زمانی مبهم، قالب زمانی نادرست

۳-۵. یکپارچه‌سازی مفاهیم

در این بخش مفاهیمی که به ویژگی‌های مشترکی اشاره دارند، به‌منظور کاهش مفاهیم و مدیریت بهتر آن‌ها با یکدیگر یکپارچه می‌شوند. بنابراین، در این مرحله چالش‌های کیفیت داده کشف شده، در گام قبلی با استفاده از ویژگی‌های مشترک و معنایی، پالایش و با یکدیگر ترکیب شده‌اند. از طرفی، یک چارچوب مناسب باید شامل راهکارهایی برای حل مشکلات کیفیت داده باشد. تجمیع مشکلات کیفیت داده، امکان

1. size and complexity

پیشنهاد فعالیت‌ها و گام‌هایی را که برای پاک‌سازی و رفع آن‌ها لازم است، آسان می‌کند. بدین‌گونه در این مرحله وظایف پاک‌سازی داده‌ها[1] برای کاهش اثر چالش‌های کیفیت داده یکپارچه شده، بر اساس ویژگی‌های آن‌ها پیشنهاد می‌شود. به بیانی دیگر، برای هر کدام از مشکلات کیفیت داده، یک رهیافت پاک‌سازی تعیین شده است. برای دستیابی به هدف فوق، چهار گام اصلی مشخص شده که شامل فعالیت‌های: «ترمیم»، «کشف ناهنجاری»، «پالایش» و «کاهش ابعاد» هستند. در جدول ۴، چالش‌های کیفیت داده، چالش کیفیت دادهٔ یکپارچه‌شدهٔ نهایی، عنوان و هدف وظیفهٔ پاک‌سازی مربوط به هر کدام از آن‌ها مشخص شده است.

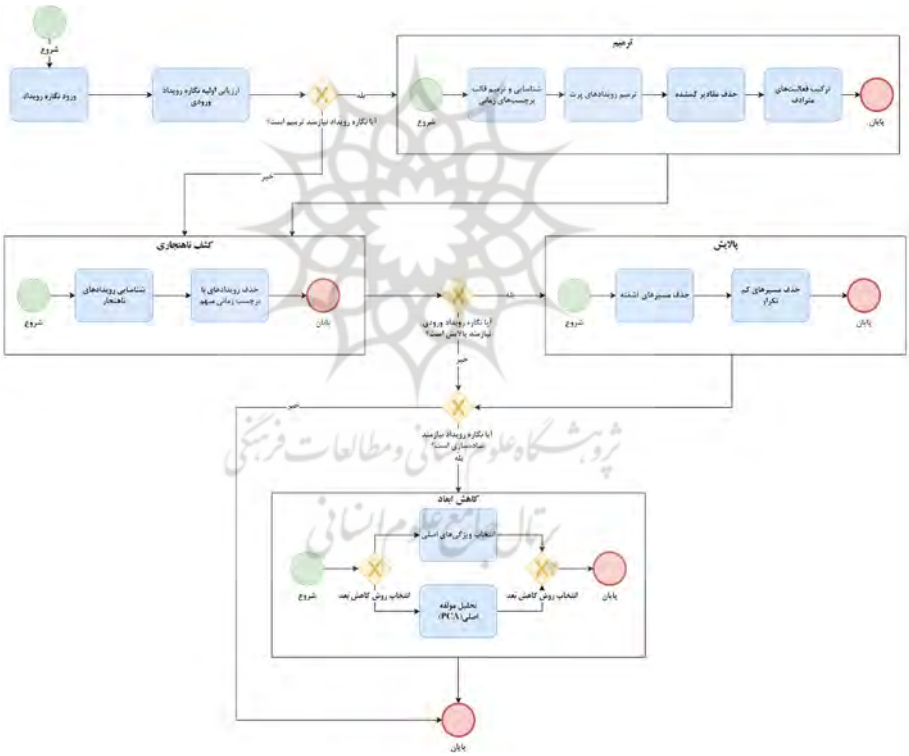
جدول ٤. چالش‌های کیفیت داده یکپارچه‌شده به‌همراه وظایف پاکسازی متناظر با آن‌ها

چالش‌های کیفیت داده	چالش کیفیت داده یکپارچه‌شده	عنوان وظیفهٔ پاک‌سازی	هدف وظیفهٔ پاکسازی
مسیرهای آشفته/ کم‌تکرار	مسیرهای آشفته/ کم‌تکرار	پالایش	حذف مسیرها و دنباله‌هایی که ممکن است باعث افزایش پیچیدگی مدل فرایند کشف‌شده توسط روش‌های کشف فرایند شود.
رویدادهای بدون برچسب رویدادهای گمشده ویژگی‌های گمشده	مقادیر گمشده	ترمیم	بهبود ساختار و کیفیت رویدادهای ذخیره‌شده درون نگارهٔ رویداد از طریق ترمیم یا حذف مقادیر گمشده
فعالیت‌های بی‌نظم، فعالیت‌های پراکنده	رویدادهای پرت	ترمیم	بهبود و جایگزینی رویدادهای با فعالیت‌های پرت و نامشخص به‌منظور افزایش کیفیت داده‌های ورودی
رویدادهای آشفته/ کم‌تکرار	رویدادهای ناهنجار	کشف ناهنجاری	شناسایی رویدادهایی که ممکن است بعد از بازسازی رویدادها در مرحلهٔ ترمیم دچار ناهنجاری شوند
قالب زمانی نادرست	قالب زمانی نادرست	ترمیم	بازسازی ترتیب رویدادها با استفاده از بازسازی ساختار قالب‌های زمانی
برچسب‌های زمانی مبهم/ نادرست	برچسب‌های زمانی مبهم	کشف ناهنجاری	یافتن برچسب‌های زمانی مبهم با هدف افزایش کیفیت مدل فرایند استخراج‌شده توسط الگوریتم‌های کشف فرایند
فعالیت‌های تکراری فعالیت‌های مترادف	فعالیت‌های مترادف	ترمیم	ترکیب فعالیت‌هایی که از دیدگاه معنا یا کارکرد با یکدگیر یکسان هستند درون یک فعالیت واحد
اندازه و پیچیدگی	اندازه و پیچیدگی	کاهش ابعاد	کاهش ابعاد داده‌های ورودی با هدف بهبود کارایی روش‌های کشف فرایند

1. data cleaning tasks

۳-۶. ترکیب مفاهیم در داخل چارچوب

در این مرحله، چارچوب مفهومی از طریق ترکیب مفاهیم به‌دست‌آمده از مرحلهٔ قبل در یک ساختار مشخص و بامعنا ایجاد می‌شود. در حقیقت، در این گام شبکهٔ معنایی بین چالش‌های کیفیت داده و وظایف پاکسازی متناظر با هر کدام از آن‌ها مشخص شده است. بدین‌سان بر اساس فعالیت‌های چهارگانهٔ شناسایی‌شده در گام پنجم، «چارچوب مفهومی برای پیش‌پردازش نگاره‌های رویداد»[1] یا به اختصار 'CF4ELP' مطابق شکل ۴، ایجاد گردید. شایان توجه است که هر فعالیت پاک‌سازی داخل چارچوب به‌صورت یک زیرفرایند پیشنهاد شده است؛ زیرا هر مرحلهٔ پاک‌سازی داده‌ها می‌تواند شامل زیرگام‌های مجزای دیگری باشد که در ادامه، به فعالیت‌های هر بخش اشاره خواهد شد.



شکل ٤. چارچوب مفهومی به همراه تشریح فعالیت‌های مربوط به هر زیرفرایند

1. conceptual framework for event log preprocessing

- **ورود نگارهٔ رویداد:** نقطهٔ آغازین هر تحلیل داده-محور، ورود داده‌ها برای انجام عملیات تحلیل است. در اینجا منظور از ورودی، یک نگارهٔ رویداد با قالب استاندارد است. قالب‌های رایج در فرایندکاوی شامل csv[1]، xes[2] هستند که چارچوب پیشنهادی با هر دو قالب اشاره‌شده به‌درستی سازگار است؛
- **ارزیابی اولیهٔ نگارهٔ رویداد ورودی:** در مرحلهٔ بعد، ارزیابی نگارهٔ رویداد به جهت ایجاد درک اولیه و ایجاد امکان تصمیم‌گیری مناسب ضروری است. در این بخش تعداد کل رویدادهای یک نگارهٔ رویداد، تعداد ویژگی‌ها، تعداد موردها و تعداد رویدادهای مجزا اندازه‌گیری می‌شود؛
- **ترمیم:** اولین بخش برای بهبود کیفیت نگارهٔ رویداد زیرفرایند ترمیم است. در گام بعد، یک دروازهٔ تصمیم وجود دارد. در این مرحله با توجه به ارزیابی‌های صورت‌گرفته در مرحلهٔ قبل بررسی می‌شود که آیا نگارهٔ رویداد نیازمند ترمیم است یا خیر؟ اگر پاسخ مثبت باشد، عملیات ترمیم انجام می‌پذیرد؛ در غیر این صورت جریان فرایند پاک‌سازی به سمت مرحلهٔ بعدی ادامه پیدا می‌کند. در اولین فعالیت متعلق به این زیرفرایند، برچسب‌های زمانی نادرست از جنبهٔ درست بودن قالب آن‌ها، تصحیح می‌شوند. در پاره‌ای از موارد برچسب‌های زمانی برای رویداد در قالب‌های متفاوتی ذخیره می‌شوند که لازم است به یک ساختار مشخص و یکسان تبدیل گردند تا بتوان بر مبنای آن‌ها فعالیت‌های یک فرایند را مرتب نمود. دومین فعالیت شامل ترمیم رویدادهای پرت است. در این بخش دنباله‌های موجود در یک نگارهٔ رویداد بررسی و اگر در یک دنباله یا مسیر، فعالیتی موجود باشد که دارای گمشدگی یا مقدار پرت است، با استفاده از محاسبهٔ تشابه دنباله با نزدیک‌ترین دنبالهٔ خود، آن فعالیت با مقدار متناظر در دنبالهٔ مشابه جایگزین می‌گردد. برای این کار از روش معرفی‌شده در مقاله Sani, van Zelst and van der Aalst (2018b) استفاده شده است؛ در گام بعد، نگارهٔ ورودی از دیدگاه وجود مقادیر گمشده در سطح ویژگی‌های مورد، فعالیت و برچسب زمانی مورد بررسی قرار می‌گیرد. در صورت وجود گمشدگی برای یک رویداد، آن رویداد با استفاده از رویکرد حذف از مجموعه داده‌ها حذف می‌گردد (Aljuaid and Sasi 2016). چهارمین وظیفهٔ این زیرفرایند به ترکیب و یکپارچه

1. comma-separated values 2. http://www.xes-standard.org/

نمودن فعالیت‌های مترادف یا یکسان که ممکن است در یک نگارهٔ رویداد موجود باشند، می‌پردازد. به‌عنوان نمونه، در یک نگارهٔ رویداد ممکن است دو عبارت پرداخت و پرداخت الکترونیکی وجود داشته باشد که هر دو به یک فعالیت اشاره دارند. پس می‌توان آن‌ها را با یکدیگر ترکیب نمود.

- **کشف ناهنجاری:** به‌هنگام انجام زیرفرایند ترمیم ممکن است بعضی از مقادیر به‌گونه‌ای اصلاح شوند که با منطق فرایند کسب‌وکار هماهنگ نیستند. پس فعالیت کشف ناهنجاری باید پس از عمل ترمیم گنجانده شود. در گام اول با استفاده از روش خوشه‌بندی، همهٔ دنبالهٔ فعالیت‌های موجود در نگارهٔ رویداد بررسی شده و به دو خوشهٔ مجزا تقسیم می‌شوند. روش خوشه‌بندی بر پایهٔ تعداد تکرار فعالیت‌ها (بسامد) عمل می‌نماید و حداقل اندازه یک خوشه ۰/۲۰ یعنی ۲۰ درصد کل دنباله‌ها در نظر گرفته شده است. معیار توقف هم تقسیم دنباله‌ها به‌عنوان ۲ خوشهٔ مجزا بوده و رویدادهایی که به خوشه با کمترین تناسب، نگاشت شده‌اند، به‌عنوان نقطهٔ پرت در نظر گرفته شده و از مجموعهٔ داده‌ها حذف خواهند شد. در گام دوم، رویدادهای بدون برچسب زمانی یا برچسب زمانی نادرست که توسط زیرفرایند ترمیم اصلاح نشده‌اند، از داده‌های رویداد ورودی کنار گذاشته خواهند شد؛
- **پالایش:** در ادامه، نقطهٔ تصمیم دیگری وجود دارد و آن اینکه آیا نگارهٔ رویداد ورودی به فعالیت پالایش نیازمند است یا خیر؟ اگر پاسخ مثبت باشد، جریان فرایند پاک‌سازی وارد زیرفرایند پالایش خواهد شد. در این گام دو وظیفهٔ اصلی وجود دارد. در وظیفهٔ اول، فعالیت‌های آشفته یا بی‌نظم که به‌صورت موردی یا بدون نظم خاصی در یک فرایند رخ می‌دهد، پالایش می‌شوند تا دقت دنبالهٔ فعالیت‌ها بیشتر گردد. برای این کار از شیوهٔ معرفی‌شده در پژوهش Tax, Sidorova and van der Aalst 2019) استفاده می‌شود. در مرحلهٔ بعد، فعالیت‌هایی که تعداد رخداد آن‌ها در فرایند کسب‌وکار کم است، از عملیات کشف فرایند حذف خواهند شد. برای این منظور، تعداد رخداد هر فعالیت در هر دنبالهٔ فعالیت (مسیر فرایند) محاسبه شده و بر اساس تعداد کل رخدادها، احتمال وقوع هر رویداد تعیین می‌شود. فعالیت‌هایی که احتمال رخداد آن‌ها از میزان آستانه کمتر باشد، از داده‌های ورودی حذف می‌شوند. حد آستانه برای این کار ۰/۹۰ تعیین شده است؛

◇ **کاهش ابعاد:** در گام بعدی نقطهٔ تصمیم‌گیری، کاهش ابعاد یا ساده‌سازی قرار دارد و تعیین می‌کند که آیا لاگ به ساده‌سازی نیازمند است یا خیر؟ این مرحله برای همهٔ نگاره‌های رویداد ضروری نیست. هدف این مرحله کاهش پیچیدگی رویدادهای ورودی به الگوریتم‌های کشف فرایند است. این عمل باعث افزایش بازدهی عملکرد روش‌های کشف فرایند نیز خواهد شد. در چارچوب پیشنهادی برای انجام کشف فرایند فقط ویژگی‌های شناسهٔ مورد، فعالیت و برچسب زمانی در نظر گرفته شده‌اند و بر این اساس می‌توان دیگر ویژگی‌های یک نگارهٔ رویداد را از عملیات کشف فرایند حذف نمود. در این بخش می‌توان از روش‌های کاهش بُعد مانند تحلیل مؤلفهٔ اساسی[1] هم استفاده نمود تا فقط ویژگی‌های با «وردایی»[2] بالا در داده‌های ورودی باقی بمانند.

۳-۷. اعتبارسنجی چارچوب مفهومی

در این گام باید چارچوب را از منظر معتبر بودن مورد سنجش قرار داد. برای سنجش چارچوب 'CF4ELP' از مجموعه داده‌های استاندارد که استفاده از آن‌ها در ادبیات فرایندکاوی رایج است، استفاده شده که نتایج حاصل از پیش‌پردازش داده‌های ورودی به روش‌های کشف فرایند با استفاده از چارچوب پیشنهادی در بخش ۴، مورد بحث و بررسی قرار گرفته است.

٤. یافته‌های پژوهش

در این بخش نتایج ارزیابی چارچوب مفهومی پیشنهادی ارائه خواهد شد. برای پیاده‌سازی چارچوب مفهومی از یک ابزار و یک کتابخانه استفاده شده است. از برنامهٔ "ProM" برای اعمال فعالیت‌های مربوط به زیرفرایندهای ترمیم، کشف ناهنجاری و پالایش استفاده شده است. همچنین برای بررسی نتایج عملکرد روش‌های کشف فرایند بر روی مجموعهٔ داده‌ها در دو حالت بدون پیش‌پردازش و پس از پیش‌پردازش، از کتابخانهٔ "PM4py" بهره برده شده است. برای پیاده‌سازی و آزمایش چارچوب پیشنهادی، از یک رایانهٔ رومیزی مجهز به پردازنده "AMD Ryzen 7 4800H" با سرعت ۲/۹ گیگاهرتز که دارای ۱۶ گیگابایت حافظهٔ اصلی است، استفاده شده است.

1. principal component analysis (PCA)
2. variance

٤-١. مجموعه داده‌های مورد استفاده

برای بررسی عملکرد چارچوب معرفی‌شده از مجموعه نگاره‌های رویداد استانداردی که برگرفته از رویدادهای مربوط به فرایندهای کسب‌وکار واقعی بوده و توسط جامعهٔ فرایندکاوی گردآوری شده‌اند، استفاده شده است. در ادبیات فرایندکاوی از این مجموعه داده‌ها برای بررسی و آزمون عملکرد الگوریتم‌ها و روش‌های فرایندکاوی استفاده می‌شود (Lopes and Ferreira 2019). در جدول ۵، نگاره‌های رویداد مورد استفاده به همراه ویژگی‌های آن‌ها آمده است.

جدول ۵. مجموعه داده‌های مورد استفاده

نگارهٔ رویداد	حوزه	تعداد رویدادها	تعداد موردها	تعداد رویدادهای مجزا	منبع
BPIC 2012	مالی و اعتباری	۲۶۲۲۰۰	۱۳۰۸۷	۳۶	Van Dongen (2012)
Hospital_Billing	سلامت	۱۵۰۲۹۱	۱۱۴۳	۶۲۴	Mannhardt (2017
Road	راه و ترابری	۵۶۱۴۷۰	۱۵۰۳۷۰	۱۱	De Leoni and Mannhardt (2015)
Sepsis	سلامت	۱۵۲۱۴	۱۰۵۰	۱۶	Mannhardt (2016)

٤-٢. معیارهای ارزیابی

با توجه به اینکه چارچوب پیشنهادی در این مطالعه بر روی پیش‌پردازش نگارهٔ رویداد با هدف بالا بردن کیفیت مدل‌های فرایندی استخراجی تمرکز دارد، به همین دلیل، از معیارهای مرتبط با اندازه‌گیری کیفیت یک مدل استخراج‌شده توسط روش‌های کشف فرایند استفاده شده است. در این مطالعه از سه بُعد[1] زیر برای این منظور بهره گرفته شده، که به‌شرح زیر است:

- تناسب[2]: رفتار ثبت‌شده در نگارهٔ رویداد در مدل فرایند استخراج‌شده به چه میزان دیده می‌شود. مقدار این معیار عددی بین ۰ تا ۱ خواهد بود. هر اندازه میزان این معیار به ۱ نزدیک‌تر باشد، تناسب مدل با نگارهٔ رویداد بیشتر است. برای محاسبهٔ این معیار از روش معرفی‌شده در مطالعهٔ Berti and van der Aalst (2019) استفاده شده است؛
- دقت[3]: رفتار شرح داده‌شده توسط مدل فرایند در نگارهٔ رویداد به چه اندازه مشاهده

1. dimension
2. fitness
3. precision

می‌شود. ارزش این معیار هم عددی بین ۰ تا ۱ است. به‌منظور اندازه‌گیری این معیار، از روش معرفی‌شده در مقالهٔ Munoz-Gama and Carmona (2010) استفاده شده است؛

◇ معیار F: ترکیبی از ابعاد تناسب و دقت در یک معیار واحد است که با استفاده از رابطهٔ ۱، محاسبه می‌شود (De Weerdt et al. 2011):

$$F - measure = (2 * Precision \times Fitness)/(Precision + Fitness) \quad (۱$$

باید توجه داشت که در بسیاری از کاربردها معیار تناسب از اهمیت بیشتری برخوردار است (Sani, van Zelst, and van der Aalst 2018b). بنابراین، در کاربردهای واقعی، مدل‌های فرایندی کشف‌شده‌ای مناسب هستند که معیار تناسب آن‌ها بزرگ‌تر از ۰/۹۰ باشد. همچنین باید بیان کرد که با بررسی ادبیات حد آستانهٔ مناسب برای معیار F در روش‌های کشف فرایند پیشنهاد نشده است.

٤-۳. نتایج ارزیابی

برای بررسی سنجش اثربخشی چارچوب مفهومی معرفی‌شده، از الگوریتم‌های رایج فرایندکاوی آلفا[1] (Van der Aalst, Weijters, and Maruster 2004)، آلفا پلاس[2] (De Medeiros et al. 2004)، هیوریستیک ماینر[3] (Weijters, van Der Aalst and De Medeiros 2006) و اینداکتیو ماینر[4] (Leemans, Fahland and van der Aalst 2013) استفاده شده و داده‌های ورودی در دو حالت بدون پیش‌پردازش و پس از پیش‌پردازش به آن‌ها اعمال گردیده‌اند. برای هر کدام از مجموعه داده‌های رویداد فوق کیفیت مدل استخراجی به‌تفکیک الگوریتم استفاده‌شده بر مبنای معیارهای بخش ۴-۲ گزارش شده است. پس از ورود مجموعه داده‌ها به الگوریتم‌های کشف فرایند، در مرحلهٔ دوم نوبت به انجام عملیات پیش‌پردازش می‌رسد. مطابق چارچوب پیشنهادی، گام‌های لازم برای انجام پیش‌پردازش برای داده‌های ورودی انجام پذیرفته و این بار داده‌های پردازش‌شده به‌عنوان دادهٔ ورودی به چهار روش کشف فرایند اشاره‌شده اعمال شده است. در جدول ۶، نتایج حاصل از معیارهای کیفیت کشف‌شده از چهار مجموعه دادهٔ معرفی‌شده در بخش ۴-۱ در دو حالت قبل و بعد از پیش‌پردازش گزارش شده است.

1. Alpha (α)
2. Alpha plus (α+)
3. heuristic miner (HM)
4. inductive miner (IM)

جدول ٦. معیارهای ارزیابی مدل فرایند کشف‌شده قبل و بعد از پیش‌پردازش بر اساس چارچوب مفهومی پیشنهادی

نگارهٔ رویداد	قبل از پیش‌پردازش				بعد از پیش‌پردازش			
	الگوریتم کشف فرایند	تناسب	دقت	معیار F	تناسب	دقت	معیار F
BPIC 2012	*α*	۰/۷۷	۰/۱۰	۰/۱۸	۰/۵۰	۰/۲۲	۰/۳۱
	α+	۰/۷۰	۰/۱۶	۰/۲۶	۰/۶۷	۰/۲۳	۰/۳۴
	HM	۰/۹۹	۰/۴۵	۰/۶۲	۰/۹۳	۰/۸۵	۰/۸۹
	IM	۱/۰۰	۰/۱۱	۰/۲۰	۱/۰۰	۰/۳۴	۰/۵۱
Hospital Billing	*α*	۰/۷۴	۰/۱۰	۰/۵۰	۰/۷۳	۰/۵۳	۰/۶۲
	α+	۰/۳۴	۰/۳۶	۰/۳۵	۰/۹۰	۰/۴۱	۰/۵۷
	HM	۰/۹۲	۰/۹۹	۰/۹۶	۰/۹۹	۰/۹۹	۰/۹۹
	IM	۱/۰۰	۰/۴۹	۰/۶۶	۱/۰۰	۰/۵۸	۰/۷۴
Road	*α*	۰/۶۷	۰/۶۶	۰/۶۶	۰/۷۶	۰/۶۵	۰/۷۰
	α+	۰/۳۸	۰/۶۶	۰/۴۸	۰/۳۸	۰/۷۵	۰/۵۰
	HM	۰/۹۲	۰/۹۹	۰/۹۵	۰/۹۸	۰/۹۹	۰/۹۹
	IM	۱/۰۰	۰/۶۹	۰/۸۲	۱/۰۰	۰/۶۸	۰/۸۱
Sepsis	*α*	۰/۲۶	۰/۴۴	۰/۳۵	۰/۳۳	۰/۸۰	۰/۴۶
	α+	۰/۷۸	۰/۳۱	۰/۴۶	۰/۶۴	۰/۹۷	۰/۷۷
	HM	۰/۹۲	۰/۷۱	۰/۸۰	۰/۹۰	۰/۷۹	۰/۸۵
	IM	۱/۰۰	۰/۱۹	۰/۳۲	۱/۰۰	۰/۴۵	۰/۶۲

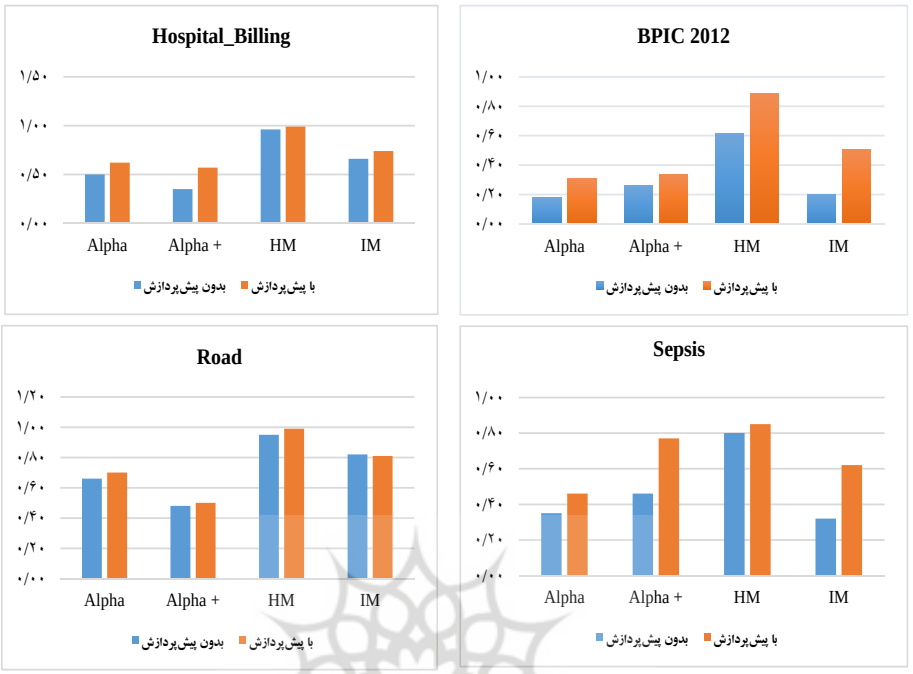
٤-٤. بحث و تبیین نتایج

همان‌طور که در جدول ۶، مشخص شده، الگوریتم α و +α به‌هنگام مواجهه با داده‌هایی که به‌صورت خام به آن‌ها وارد می‌شود، با مشکل روبه‌رو خواهند شد. آشکار است که دو الگوریتم فوق در معیار دقت نتایج مناسبی تولید نمی‌کنند و نخواهند توانست نگارهٔ رویداد را بازتولید نمایند. به همین سبب، معیار F برای الگوریتم‌های فوق در وضعیت مناسبی قرار نخواهد داشت و به همین دلیل، برای دستیابی به نتیجهٔ بهتر باید

نگاره‌های رویداد را خلاصه نمود تا رفتارهای پیچیدهٔ آن‌ها کاهش یافته و آن‌ها را به این الگوریتم‌ها اعمال نمود. از طرفی، اصلاح برچسب‌های زمانی منجر به تفسیر بهتر ترتیب فعالیت‌ها می‌شود و به افزایش معیار دقت مدل خروجی حاصل از الگوریتم‌های فوق کمک خواهد نمود.

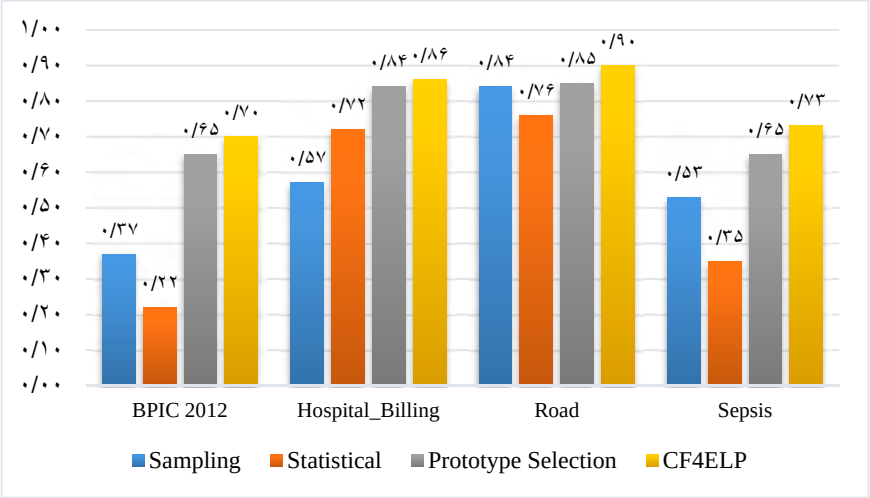
این نکته در مورد الگوریتم‌های پیشرفته‌تر که در دنیای واقعی از آن‌ها بیشتر استفاده می‌شود، هم معتبر است. الگوریتم HM میزان تکرار فعالیت‌های ثبت‌شده در یک نگارهٔ رویداد را در نظر گرفته و از قابلیت مواجهه با حلقه‌های کوتاه نیز برخوردار است. این الگوریتم در بین همهٔ روش‌ها در معیار دقت بهترین نتیجه را کسب نموده است. الگوریتم IM توانایی درک فعالیت‌هایی را که به‌صورت ضمنی در داده‌های رویداد یک فرایند کسب‌وکار رخ می‌دهند، دارد و در بیشتر موارد مدل‌های کسب‌وکار کامل و صحیحی را به‌عنوان خروجی نتیجه می‌دهد. این الگوریتم در به‌دست آوردن تناسب بالا موفق عمل می‌نماید، اما در معیار دقت نسبت به الگوریتم HM ضعیف‌تر است. در مجموع، نتایج به‌دست‌آمده نشان داد که روش HM در معیار F از دیگر روش‌ها، از جمله الگوریتم IM موفق‌تر بوده، و این امر به این دلیل است که روش HM مدل‌های فرایند ساده‌تری را استخراج می‌نماید.

با مقایسهٔ نتایج گزارش‌شده در جدول ۶، نتیجه گرفته می‌شود که به هر میزان بتوان با استفاده از عملیات پیش‌پردازش، پیچیدگی داده‌های رویداد را کاهش داد، نتایج حاصل از روش‌های HM و IM در معیار دقت بهبود بیشتری به‌دست می‌آورند. این مهم در شکل ۵، به تصویر درآمده است. می‌توان نتیجه گرفت که الگوریتم HM در مجموعه داده‌های ساده‌تر بهترین عملکرد را از خود نشان می‌دهد، اما اگر پیچیدگی مجموعه دادهٔ رویداد بیشتر شود، روش IM توانایی بهتری برای متناسب شدن مدل فرایندی استخراجی با نگارهٔ رویداد خواهد داشت؛ هرچند مدل کسب‌وکار استخراجی نمی‌تواند همهٔ رفتارهای مشاهده‌شده در لاگ را بازتولید نماید. به همین دلیل، انجام پیش‌پردازش برای بهتر شدن این چالش کمک زیادی خواهد نمود که مطابق انتظار پس از پیش‌پردازش معیار F برای روش IM بهبود یافته است.



شکل ۵. نمایی از مقایسهٔ معیار F بین الگوریتم‌های مختلف در دو حالت بدون پیش‌پردازش و پردازش‌شده

برای اعتبارسنجی صحت عملکرد چارچوب پیشنهادی، نتایج حاصل از آن با دیگر روش‌های موجود در ادبیات مقایسه شده است. برای انجام ارزیابی مناسب، میانگین معیار F برای الگوریتم‌های HM و IM پس از پیش‌پردازش محاسبه شده و آنگاه، مقدار نهایی با میانگین معیار F حاصل از مدل‌های کشف‌شده توسط روش‌های «نمونه‌برداری» (Fani Sani, Zelst, and van der Aalst 2019)، «پیش‌پردازش آماری» (Bauer et al. 2018) و «انتخاب نمونهٔ اولیه» (Sani, Boltenhagen, and van der Aalst 2019) که در بخش ۲، معرفی شدند، سنجیده شده است. هر سه روش فوق پس از اعمال پیش‌پردازش بر روی نگاره‌های رویداد، آن‌ها را به الگوریتم‌های کشف فرایند اعمال، و خروجی حاصل را گزارش نموده‌اند. سه روش اشاره‌شده به‌دنبال یافتن بهترین دنبالهٔ فعالیت‌ها هستند که بتوانند افزون بر خلاصه‌سازی نگارهٔ رویداد، بهترین تقریب از رفتارهای ثبت‌شده در نگارهٔ رویداد را استخراج نمایند تا مدل حاصل بتواند در هر سه معیار تناسب، دقت و F به نتایج مطلوب دست یابد. نتایج بررسی عملکرد چارچوب پیشنهادی با سه روش پیشین بر پایهٔ معیار F در شکل ۶، مشخص شده است.



شکل ٦. مقایسهٔ نتایج چارچوب پیشنهادی با دیگر روش‌های ارائه‌شده بر مبنای معیار F

رویکرد پیشنهادی به‌دلیل در نظر گرفتن مهم‌ترین عوامل تأثیرگذار بر روی کیفیت نگارهٔ رویداد به‌صورت محسوس عملکرد بهتری نسبت به سه روش قبلی از خود نشان داده، زیرا عمل پیش‌پردازش را فقط به‌عنوان یک فعالیت با یک هدف واحد در نظر نگرفته است، بلکه این عمل را ترکیبی از چندین فعالیت پیش‌پردازش فرض نموده که انجام هر مرحلهٔ آن به افزایش معیارهای کیفیت نتایج خروجی و پالایش رفتارهای دارای اهمیت کمتر منجر می‌شود. آزمایشات نشان داد که بین ۸۰ تا ۹۰ درصد فعالیت‌هایی با بیشترین تکرار، دربرگیرندهٔ بیش از ۶۵ درصد از همه فعالیت‌های ثبت‌شده در یک نگارهٔ رویداد هستند. چارچوب پیشنهادی با حفظ کردن رویدادهای با اهمیت بیشتر، بازسازی رویدادهای پرت و یکپارچه نمودن فعالیت‌های مترادف امکان کشف مدل‌های با کیفیت بالاتری را فراهم ساخته است. از طرفی با تصحیح قالب برچسب‌های زمانی امکان مرتب نمودن بهتر فعالیت‌ها فراهم شده و مدل نهایی بازنمای درستی از ترتیب فعالیت‌ها ارائه خواهد داد. همراستا با موارد فوق، مد نظر قرار دادن ویژگی‌های مهم برای الگوریتم‌های کشف فرایند و حذف ویژگی‌های با اهمیت کمتر، افزون بر کاهش منابع مصرفی و زمان اجرا، کارایی الگوریتم‌های کشف فرایند را افزایش می‌دهد.

٤-٥. کاربردهای چارچوب پیشنهادی

کیفیت پایین داده‌ها یکی از دلایل اصلی ناکامی در پروژه‌های فرایندکاوی است.

اگـر داده‌هـای رویـداد قابـل اعتمـاد نباشـند، نتایـج فراینـدکاوی از ارزش کمتـری برخـوردار خواهنـد بـود. بـا وجـود دیدگاه‌هـای متعـدد در مـورد تأثیـر کیفیـت داده‌هـا در پژوهش‌هـای مربـوط بـه ایـن زمینـه، همـهٔ آن‌هـا در انتها در یـک نقطـه بـا یکدیگـر اشـتراک دارنـد. داده‌هـای بـا کیفیـت پاییـن می‌تواننـد هزینه‌هـای مالـی ناخواسـته‌ای را بـه سـازمان‌ها تحمیـل کننـد (Laranjeiro, Soydemir and Bernardino 2015). از ایـن رو، نیـاز بـه فرایندهـای ارزیابـی و همچنین اولویت‌بنـدی اقدامـات مناسـب بـرای بهبـود کیفیـت داده‌هـا بـرای متخصصـان و تحلیلگـران داده و کسـب‌وکار ضـروری خواهـد بـود (Loshin 2010). بایـد توجـه داشـت کـه در زمینـهٔ مدیریـت کیفیـت داده‌هـا چارچوب‌هـای مختلفـی ارائـه شـده‌اند کـه هـدف اصلـی آن‌هـا ارزیابـی کیفیـت داده‌هـای موجـود در یـک سـازمان اسـت (Wang and Strong 1996; Cichy and Rass 2019). چارچـوب پیشـنهادی در پژوهـش حاضـر بـا هـدف ارزیابـی داده‌هـای رویـداد ذخیره‌شـده در سـامانه‌های اطلاعاتـی ایجـاد شـده اسـت کـه بـا توجـه بـه شناسـایی مهم‌تریـن مشـکلات کیفیـت داده در زمینـهٔ فراینـدکاوی، امـکان بررسـی مهم‌تریـن چالش‌هـای کیفیـت داده بـرای متخصصـان را در یـک سـاختار یکپارچـه فراهـم می‌نمایـد. از طرفـی، «ایپلـر و ویتیـگ» بیـان می‌کننـد کـه یـک چارچـوب مناسـب افـزون بـر قابلیـت ارزیابـی بایـد برنامـه و طرحـی بـرای حـل مشـکلات کیفیـت داده‌هـا داشـته باشـد (Eppler and Wittig 2000). در طراحـی چارچـوب پیشـنهادی تـلاش شـده ایـن ویژگـی مـورد توجـه قـرار گیـرد؛ بـه ایـن صـورت کـه بـا پیشـنهاد و اولویت‌بنـدی زیرفرایندهـای چهارگانـه بـرای پاک‌سـازی داده‌هـای رویـداد، امـکان حـل یـا کاهـش اثـر مشـکلات کیفیـت داده را نیـز بـرای متخصصـان داده و فراینـد امکان‌پذیـر سـازد. «وین و صدیـق» معرفـی چارچوب‌هـای کیفیـت داده‌ای را در فراینـدکاوی پیشـنهاد می‌نماینـد کـه سـه ویژگـی «درک مشـترک از مشـکلات کیفیـت داده بـرای ذی‌نفعـان»، «مدل‌سـازی ترجیحـات تحلیلگـران» و «ایجـاد شـفافیت بـرای درک اثـر بهبـود کیفیـت داده‌هـا در نتایـج خروجـی» را داشـته باشـند (Wynn and Sadiq 2019). راهـکار معرفی‌شـده در ایـن پژوهـش امـکان ایجـاد درک مشـترک از طریـق دسـته‌بندی چالش‌هـای کیفیـت داده در یـک سـاختار جامـع را پدیـد مـی‌آورد. همچنیـن، بـا پشـتیبانی از شـرایط تصمیم‌گیـری، امـکان انعطـاف و در نظـر گرفتـن ترجیحـات کاربـران و تحلیلگـران را فراهـم می‌نمایـد. نتایـج اسـتفاده از چارچـوب ایجادشـده بـر روی مجموعـه داده‌هـای معرفی‌شـده در بخـش ۳-۴ به‌خوبـی نشـان‌دهندهٔ تأثیـر پیش‌پـردازش داده‌هـای ورودی بـر روی نتایـج خروجـی بـوده اسـت.

بـا توجـه بـه ویژگی‌هـای بیان‌شـده، چارچـوب معرفی‌شـده در ایـن پژوهـش می‌توانـد

به‌عنوان یک راهنما و راهبرد مشخص توسط تحلیلگران داده و کسب‌وکار مورد استفاده قرار گیرد تا با دقت عمل بیشتری به پیش‌پردازش داده‌ها به‌عنوان بخشی مهم در هر پروژهٔ فرایندکاوی اقدام نمایند. همچنین، این رهیافت باعث کاهش بینش‌های تجربی و روش‌های مبتنی بر سعی و خطا جهت افزایش کیفیت داده‌های رویداد خواهد شد.

۵. نتیجه‌گیری و پیشنهادهایی برای پژوهش‌های آینده

فرایندکاوی به‌عنوان شاخه‌ای جدید از علم داده به‌دنبال کشف و تحلیل فرایندهای کسب‌وکار از داده‌های رویداد ذخیره‌شده در سامانه‌های اطلاعاتی است. در بیشتر وقت‌ها در روش‌های کشف فرایند فرض بر این است که نگاره‌های رویداد مورد استفاده کامل، بدون خطا و مقادیر گمشده هستند. بنابراین، ممکن است انجام کشف فرایند بدون در نظر گرفتن کیفیت نگاره‌های رویداد منجر به ایجاد دیدگاه نادرست در مورد فرایندهای در حال اجرا شود. برای حل این چالش، در این پژوهش افزون بر شناسایی مهم‌ترین مشکلات کیفیت داده در فرایندکاوی، به معرفی و توسعهٔ یک چارچوب مفهومی بر پایهٔ یک روش کیفی اقدام شده است. هدف اصلی چارچوب مفهومی فوق ارائهٔ یک راهنمای مشخص به تحلیلگران برای مقابله با چالش‌های کیفیت داده در یک نگارهٔ رویداد به‌هنگام انجام فعالیت کشف فرایند است.

برای بررسی کارایی چارچوب معرفی‌شده از چهار مجموعه دادهٔ استاندارد استفاده شده که نتایج خروجی از چهار روش کشف فرایند شامل الگوریتم‌های α، α+، HM و IM بر روی داده‌های پردازش‌شده نشان داد که ابعاد تناسب، دقت و معیار F مدل کسب‌وکار استخراج‌شده، نسبت به حالت بدون پیش‌پردازش بهبود یافته است. همچنین، برای اعتبارسنجی چارچوب مفهومی پیشنهادی نتایج ارزیابی حاصل از پیش‌پردازش داده‌ها توسط چارچوب فوق با سه روش «نمونه‌برداری»، «پیش‌پردازش آماری» و «انتخاب نمونهٔ اولیه» مقایسه شده است. این مقایسه نشان داد که عملکرد چارچوب پیشنهادی نسبت به سه روش اشاره‌شده بر اساس ارزیابی معیار F بهتر بوده است.

بنابراین، با توجه به سؤالات پژوهش که در بخش ۱، عنوان شد، در پاسخ به پرسش اول باید بیان نمود که چارچوب پیشنهادی بدون استفاده از دانش مرجع و مستندات سازمانی در مورد فرایندهای کسب‌وکار توانست عمل پیش‌پردازش را انجام دهد. برای پاسخ به پرسش دوم، مهم‌ترین عوامل اثرگذار بر روی کیفیت داده‌ها پس از استخراج،

پالایش و تجمیع شناسایی شدند که شامل: «رویدادهای آشفته/ کم‌تکرار»، «رویدادهای پرت»، «رویدادهای ناهنجار»، «مقادیر گمشده»، «قالب زمانی نادرست»، «برچسب‌های زمانی مبهم»، «فعالیت‌های مترادف» و «اندازه و پیچیدگی» هستند. آنگاه، وظایف پاک‌سازی برای کاهش اثر هر کدام از چالش‌های شناسایی‌شده پیشنهاد شدند که شامل فعالیت‌های «ترمیم»، «کشف ناهنجاری»، «پالایش» و «کاهش ابعاد» هستند. با انجام فعالیت‌های پاک‌سازی اشاره‌شده، کیفیت داده‌های رویداد ورودی افزایش یافت که منجر به بهبود معیارهای کیفیت مدل‌های فرایند حاصل از روش‌های کشف فرایند شده است. در انتها، در پاسخ به پرسش سوم، چارچوب پیشنهادی توانست مهم‌ترین مشکلات کیفیت داده و وظایف پاک‌سازی متناظر با هر کدام از آن‌ها را در یک ساختار منسجم و روشن به‌صورت یک فرایند یکپارچه کنار یکدیگر قرار دهد.

نتایج این پژوهش می‌تواند به‌عنوان یک شیوه‌نامهٔ مشخص به متخصصان و تحلیلگران داده و فرایند برای حل چالش‌های کیفیت داده در پروژه‌ها و تحقیقات فرایندکاوی کمک کند. همچنین، یافته‌های این مطالعه می‌تواند چشم‌اندازهای پژوهشی جدیدی برای محققان فرایندکاوی در زمینهٔ کیفیت داده ترسیم نماید.

پژوهش حاضر شامل ۳ محدودیت اصلی است که به‌عنوان فعالیت‌های پژوهشی آینده پیشنهاد می‌شوند:

- ◇ استفاده از معیارهای متفاوت دیگر برای مقایسهٔ نتایج الگوریتم‌های کشف فرایند در فرایندکاوی می‌تواند قضاوت بهتری برای عملکرد روش‌های پیش‌پردازش باشد. معیار F با وجود ترکیب خاصیت معیارهای تناسب و دقت، به‌تنهایی بیانگر عملکرد مناسب یک الگوریتم کشف فرایند نیست. استفاده از معیارهای پیچیدگی و همچنین تأثیر پیش‌پردازش بر کارایی الگوریتم‌های کشف فرایند از جنبهٔ زمان اجرا می‌تواند برای تحقیقات آینده مد نظر قرار گیرد؛
- ◇ یافتن بهترین تنظیمات برای پیش‌پردازش به‌ویژه در یافتن نقاط پرت، پالایش و حذف فعالیت‌های آشفته یا با تکرار کم، یکی از چالش‌برانگیزترین موارد مربوط به مدیریت کیفیت داده‌های رویداد است. به‌صورت کلی، اینکه یک فعالیت در یک فرایند کمتر از دیگر فعالیت‌ها رخ می‌دهد، معیار مناسبی برای حذف آن نخواهد بود؛
- ◇ هر فعالیت پیش‌پردازش که در چارچوب پیشنهادی ارائه شده است، می‌تواند با روش‌های متفاوتی اجرا شود. به همین دلیل، می‌توان اضافه کردن رویکردهایی را

که بتوانند به‌صورت دقیق روش‌های مناسبی برای یک فعالیت پیش‌پردازش پیشنهاد دهند، به‌عنوان یک شکاف تحقیق در نظر گرفت. بر این اساس، می‌توان با استفاده از مفهوم هستان‌شناسی و درخت تصمیم ساختار چارچوب مفهومی را روشن‌تر و انعطاف‌پذیری آن را در مقابل مجموعه داده‌های مختلف افزایش داد.

فهرست منابع

دوماس، مارلون، مارچلو لازرا، جان مندلینگ، و هاجو ریجرز. ۲۰۱۳. *مبانی مدیریت فرایندهای کسب‌وکار*. ترجمه محمدحامد جعفرزاده، جلیل حیدری دهویی و سید محسن رهنمافرد. ۱۳۹۹. تهران: دانشگاه تهران، مؤسسه انتشارات.

شامی زنجانی، مهدی، فراز نبیبی و شادی ایران‌دوست. ۱۳۹۹. *ناخدایی دیجیتال: راهنمای تحول سازمان‌ها در عصر دیجیتال*. تهران: آریانا قلم.

فن در آلست، ویل. ۲۰۱۱. *فرایندکاوی: کشف، تطبیق و بهبود فرایندهای کسب‌وکار*. ترجمه سید حسین سیادت و راضیه همتی گشتاسب. ۱۳۹۴. تهران: دانشگاه شهید بهشتی، مرکز چاپ و انتشارات.

References

Aljuaid, Tahani, and Sreela Sasi. 2016. Proper imputation techniques for missing values in data sets. 2016 international conference on data science and engineering (ICDSE). IEEE.

Andrews, Robert, Suriadi Suriadi, Chun Ouyang, and Erik Poppe. 2018. Towards event log querying for data quality. OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Valletta, Malta.

Andrews, Robert, Christopher GJ van Dun, Moe Thandar Wynn, Wolfgang Kratsch, MKE Röglinger, and Arthur HM ter Hofstede. 2020. Quality-informed semi-automated event log generation for process mining. *Decision Support Systems* 132: 113265.

Augusto, Adriano, Raffaele Conforti, Marlon Dumas, Marcello La Rosa, and Artem Polyvyanyy. 2019. Split miner: automated discovery of accurate and simple business process models from event logs. *Knowledge and Information Systems* 59 (2): 251-284.

Ayo, Femi Emmanuel, Olusegun Folorunso, and Friday Thomas Ibharalu. 2017. A probabilistic approach to event log completeness. *Expert Systems with Applications* 80: 263-272.

Batyuk, A Ye, and Volodymyr V Voityshyn. 2018. Process Mining: Applied Discipline and Software Implementations. *Research Bulletin of the National Technical University of Ukraine ?* (5): 22-36.

Bauer, Martin, Arik Senderovich, Avigdor Gal, Lars Grunske, and Matthias Weidlich. 2018. How much event data is enough? A statistical framework for process discovery. International Conference on Advanced Information Systems Engineering. Tallinn, Estonia.

Berti, Alessandro, and Wil MP van der Aalst. 2019. Reviving Token-based Replay: Increasing Speed While Improving Diagnostics. Diagnostics. *19th International Conference on Application of Concurrency to System Design*, Aachen, Germany.

ATAED@ Petri Nets/ACSD.

Bose, R. P. Jagadeesh Chandra, Ronny S Mans, and Wil MP van der Aalst. 2013. Wanna improve process mining results?: it's high time we consider data quality issues seriously. *Business Process Management reports reports* 1302.

Burattin, Andrea. 2015. Process mining techniques in business environments. In *volume 207 of Lecture Notes in Business Information Processing*. Berlin: Springer.

Chen, Qifan, Yang Lu, Charmaine Tam, and Simon Poon. 2021. A Novel Approach to Detect Redundant Activity Labels For More Representative Event Logs." *arXiv preprint arXiv: 2103. 16061*.

Cichy, Corinna, and Stefan Rass. 2019. An overview of data quality frameworks. *IEEE Access* 7: 24634-24648.

Conforti, Raffaele, Marcello La Rosa, and A Ter Hofstede. 2018. *Timestamp repair for business process event logs*. Melbourne: University of Melbourne:

De Leoni, Massimiliano, and Felix Mannhardt. 2015. *Road Traffic Fine Management Process*. edited by 4TU.ResearchData. Dataset. https://doi.org/10.4121/uuid:270fd440-1057-4fb9-89a9-b699b47990f5

De Medeiros, Ana Karla A, Boudewijn F Van Dongen, Wil Van der Aalst, and AJMM Weijters. 2004. Process mining for ubiquitous mobile systems: an overview and a concrete algorithm. International Workshop on Ubiquitous Mobile Information and Collaboration Systems. Second CAiSE Workshop, UMICS 2004, Riga, Latvia, Revised Selected Papers 2 (pp. 151-165). Springer Berlin Heidelberg.

De Weerdt, Jochen, Manu De Backer, Jan Vanthienen, and Bart Baesens. 2011. A robust F-measure for evaluating discovered process models. 2011 IEEE Symposium on Computational Intelligence and Data Mining (CIDM). IEEE.

Emamjome, Fahame, Robert Andrews, Arthur ter Hofstede, and Hajo Reijers. 2020. Alohomora: Unlocking data quality causes through event log context. Proceedings of the 28th European Conference on Information Systems (ECIS2020).

Eppler, Martin J, and Dörte Wittig. 2000. Conceptualizing Information Quality: A Review of Information Quality Frameworks from the Last Ten Years. Proceedings of the 2000 Conference on Information Quality; 2000; Cam-bridge (MA), USA.

Fani Sani, Mohammadreza, Sebastiaan J van Zelst, and Wil MP van der Aalst. 2019. The impact of event log subset selection on the performance of process discovery algorithms. European Conference on Advances in Databases and Information Systems. ADBIS 2019 Short Papers, Workshops BBIGAP, QAUCA, SemBDM, SIMPDA, M2P, MADEISD, and Doctoral Consortium, Bled, Slovenia, September 8–11, 2019, Proceedings 23 (pp. 391-404). Springer International Publishing.

Ferreira, Diogo R, and Daniel Gillblad. 2009. Discovering process models from unlabelled event logs. International Conference on Business Process Management. Ulm, Germany

Ghionna, Lucantonio, Gianluigi Greco, Antonella Guzzo, and Luigi Pontieri. 2008. Outlier detection techniques for process mining applications. International symposium on methodologies for intelligent systems. Toronto, Canada.

Helal, Iman, and Ahmed Awad. 2020. Correlating Unlabeled Events at Runtime. *arXiv preprint arXiv:2004.09971*.

Jabareen, Yosef. 2009. Building a conceptual framework: philosophy, definitions, and procedure. *International journal of qualitative methods* 8 (4): 49-62.

Laranjeiro, Nuno, Seyma Nur Soydemir, and Jorge Bernardino. 2015. A survey on data quality: classifying poor data. 2015 IEEE 21st Pacific rim international symposium on dependable computing (PRDC). IEEE.

Leemans, Sander JJ, Dirk Fahland, and Wil MP van der Aalst. 2013. Discovering block-structured process models from event logs-a constructive approach. International conference on applications and theory of Petri nets and concurrency. Berlin Heidelberg.

Li, Feng. 2020. Leading digital transformation: three emerging approaches for managing the transition. *International Journal of Operations & Production Management* 40 (6): 809-817.

Lopes, Iezalde F, and Diogo R Ferreira. 2019. A survey of process mining competitions: the BPI challenges 2011–2018. International Conference on Business Process Management. Vienna, Austria.

Loshin, David. 2010. *The practitioner's guide to data quality improvement* Burlington: Morgan Kaufmann.

Lu, Ke, Xianwen Fang, Na Fang, and Esther Asare. 2021. Discovery of effective infrequent sequences based on maximum probability path. *Connection Science* 34 (1): 63-82.

Lu, Xixi, Dirk Fahland, Frank van den Biggelaar, and Wil van der Aalst. 2016. Handling duplicated tasks in process discovery by refining event labels. International Conference on Business Process Management. Rio de Janeiro, Brazil

.Lu, Xixi, Dirk Fahland, and Wil MP van der Aalst. 2016. In Proceedings of the BPM Demo Track 2016 Co-located with the 14th International Conference on Business Process Management (BPM 2016), Rio de Janeiro, Brazil, 21 September 2016; pp. 44–49.

Ly, Linh Thao, Conrad Indiono, Jürgen Mangler, and Stefanie Rinderle-Ma. 2012."Data transformation and semantic log purging for process mining. International Conference on Advanced Information Systems Engineering. Gdansk, Poland.

Mannhardt, Felix. 2016. *Sepsis Cases - Event Log*. edited by 4TU.ResearchData. https://doi.org/10.4121/uuid:915d2bfb-7e84-49ad-a286-dc35f063a460

Mannhardt, Felix. 2017. *Hospital Billing - Event Log*. edited by 4TU.ResearchData. https://doi.org/10.4121/uuid:76c46b83-c930-4798-a1c9-4be94dfeb741

Munoz-Gama, Jorge, and Josep Carmona. 2010. A fresh look at precision in process conformance. International Conference on Business Process Management. Hoboken, NJ, USA. Pyle, Dorian. 1999. *Data preparation for data mining*. San Fransisco: Morgan Kaufmann.

Rocco, Tonette S, and Maria S Plakhotnik. 2009. Literature reviews, conceptual frameworks, and theoretical frameworks: Terms, functions, and distinctions. *Human Resource Development Review* 8 (1):120-130.

Sadeghianasl, Sareh, Arthur HM ter Hofstede, Moe T Wynn, and Suriadi Suriadi. 2019. A contextual approach to detecting synonymous and polluted activity labels in process event logs. OTM Confederated International Conferences On the Move to Meaningful Internet Systems. Rhodes, Greece.

Sani, Mohammadreza Fani. 2020. Preprocessing Event Data in Process Mining. CAiSE (Doctoral Consortium). Grenoble, France.

_____, Mathilde Boltenhagen, and Wil van der Aalst. 2019. Prototype selection based on clustering and conformance metrics for model discovery. *arXiv preprint arXiv:1912.00736*.

Sani, Mohammadreza Fani, Sebastiaan J van Zelst, and Wil MP van der Aalst. 2017. Improving process discovery results by filtering outliers using conditional behavioural probabilities. International Conference on Business Process Management.

_____. 2018a. Applying sequence mining for outlier detection in process mining. OTM Confederated International Conferences" On the Move to Meaningful Internet Systems. Barcelona, Spain.

_____. 2018b. Repairing outlier behaviour in event logs. International Conference on Business Information Systems. Berlin, Germany.

Sim, Sunghyun, Hyerim Bae, and Yulim Choi. 2019. Likelihood-based multiple imputation by event chain methodology for repair of imperfect event logs with missing data. 2019 International Conference on Process Mining (ICPM). IEEE.

Succar, Bilal. 2009. Building information modelling framework: A research and delivery foundation for industry stakeholders. *Automation in construction* 18 (3):357-375.

Suriadi, Suriadi, Robert Andrews, Arthur HM ter Hofstede, and Moe Thandar Wynn. 2017. Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. *Information systems* 64: 132-150.

Tax, Niek, Natalia Sidorova, and Wil MP van der Aalst. 2019. Discovering more precise process models from event logs by filtering out chaotic activities. *Journal of Intelligent Information Systems* 52 (1): 107-139.

Van Der Aalst, Wil. 2016. *Process Mining, Data science in action*, *Process mining*. Heidelberg: Springer.

_____, Arya Adriansyah, Ana Karla Alves De Medeiros, Franco Arcieri, Thomas Baier, Tobias Blickle, Jagadeesh Chandra Bose, Peter Van Den Brand, Ronald Brandtjen, and Joos Buijs. 2011. Process mining manifesto. International conference on business process management. Berlin Heidelberg.

Van der Aalst, Wil, Ton Weijters, and Laura Maruster. 2004. Workflow mining: Discovering process models from event logs. *IEEE transactions on knowledge and data engineering* 16 (9): 1128-1142.

Van Dongen, Boudewijn. 2012. *BPI Challenge 2012, Event log of a loan application process*. edited by 4TU.ResearchData. https://doi.org/10.4121/uuid:3926db30-f712-4394-aebc-75976070e91f

Verhulst, Rick. 2016. Evaluating quality of event data within event logs: an extensible framework. Master's thesis, Eindhoven University of Technology.

Wang, Richard Y, and Diane M Strong. 1996. Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12 (4): 5-33.

Weijters, AJMM, Wil MP van Der Aalst, and AK Alves De Medeiros. 2006. Process mining with the heuristics miner-algorithm. *Technische Universiteit Eindhoven, Tech. Rep. WP* 166: 1-34.

Wynn, Moe Thandar, and Shazia Sadiq. 2019. Responsible process mining-a data quality perspective. International Conference on Business Process Management. Vienna, Austria.

Yi, Guo, and Zhang Peng. 2019. Novel Approach to Discover Precise Process Model by Filtering out Log Chaotic Activities. *Journal of Computers* 30 (4): 140-150.

Zamora-Polo, Francisco, Amalia Luque Sendra, Francisco Aguayo-Gonzalez, and Jesus Sanchez-Martin. 2019. Conceptual framework for the use of building information modeling in engineering education. *International Journal of Engineering Education* 35 (3): 744-755.

احمد صالحی

متولد ۱۳۶۸، دارای مدرک کارشناسی ارشد در رشتهٔ مهندسی فناوری اطلاعات از دانشگاه علم وصنعت ایران است. ایشان هم‌اکنون دانشجوی دکتری مهندسی فناوری اطلاعات در دانشگاه تربیت مدرس است. تحلیل و طراحی سیستم‌های اطلاعاتی، مدیریت فرایندهای کسب‌وکار، فرایندکاوی و یادگیری ماشین از جمله علایق پژوهشی وی است.

محمد اقدسی

دارای مدرک دکتری مهندسی صنایع از دانشگاه تسوکوبا ژاپن است. ایشان هم‌اکنون استاد دانشکدهٔ مهندسی صنایع وسیستم‌های دانشگاه تربیت مدرس است.

مدیریت فرایندهای کسب‌وکار، مهندسی مجدد فرایندها، مدیریت تحولات سازمانی، طراحی سازمان، مدیریت دانش فرایندمحور و سیستم‌های مدیریت دانش از جمله علایق پژوهشی وی است.

توکتم خطیبی

دارای مدرک دکتری مهندسی صنایع از دانشگاه تربیت مدرس است. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی صنایع و سیستم‌های دانشگاه تربیت مدرس است.

یادگیری ماشین، تحلیل دادگان و پردازش تصاویر پزشکی از جمله علایق پژوهشی وی است.

مجید شیخ‌محمدی

متولد ۱۳۵۱، دارای مدرک دکتری و پسادکتری در رشتهٔ مهندسی طراحی سیستم‌ها از دانشگاه واترلو کانادا است. ایشان هم‌اکنون دانشیار دانشکدهٔ مهندسی صنایع و سیستم‌های دانشگاه تربیت مدرس است.

تصمیم‌گیری استراتژیک و کاربرد نظریهٔ بازی‌ها در تحلیل مناقشات از جمله علایق پژوهشی وی است.