

مروری بر روش‌های مبتنی بر یادگیری ماشین در تولید لاگ رخداد مصنوعی

جلال رضایی نور، منصوره یاری ایلی

دکترای مهندسی صنایع، استاد، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه قم

Email: j.rezaee@qom.ac.ir

دکترای مهندسی فناوری اطلاعات، پژوهشگر پسادکتری، گروه مهندسی صنایع، دانشکده فنی و مهندسی، دانشگاه قم

Email: mansureyari@gmail.com

چکیده

مقدمه: اخیراً فرایند کاوی به‌عنوان یک حوزه نوظهور از علم داده نقش مهمی در استخراج دانش از داده‌های ثبت‌شده در سیستم‌های اطلاعاتی ایفا کرده است. هسته اصلی هر پروژه فرایند کاوی، لاگ رخداد است. هرچند کمبود داده‌های واقعی به‌دلیل محدودیت‌های حریم خصوصی، هزینه‌های جمع‌آوری و کیفیت پایین داده، توسعه الگوریتم‌های فرایند کاوی را با چالش مواجه کرده است. تولید داده‌های مصنوعی که ضمن حفظ ویژگی‌های آماری و ساختاری فرایند، ارتباط مستقیمی با نمونه‌های واقعی نداشته باشند راهکاری مؤثر در این زمینه است. هدف این پژوهش ارائه مروری نظام‌مند بر روش‌های تولید لاگ رخداد مصنوعی است که بر پایه الگوریتم‌های یادگیری ماشین توسعه یافته‌اند.

روش شناسی: پژوهش حاضر براساس چارچوب مرور نظام‌مند ارائه‌شده توسط (Xiao and Watson, 2019) انجام شده است. این چارچوب شامل تدوین پروتکل پژوهش، تعریف پرسش‌های تحقیق، جستجوی نظام‌مند منابع، غربالگری مطالعات، استخراج داده‌ها و تحلیل یافته‌ها است. در این مطالعه، ابتدا پرسش‌های پژوهش تدوین شد و سپس فرآیند جستجو، انتخاب و تحلیل مقالات مطابق با پروتکل مرور نظام‌مند اجرا گردید. براین اساس پرسش‌ها شامل (۱) بررسی روش‌ها، الگوریتم‌ها و رویکردهای مبتنی بر یادگیری ماشین برای تولید لاگ رخداد مصنوعی در فرایند کاوی، (۲) مهم‌ترین کاربردهای لاگ رخداد مصنوعی در مطالعات فرایند کاوی، و (۳) روش‌های ارزیابی کیفیت، سودمندی و کارایی لاگ رخداد مصنوعی تعریف شدند. جستجوی نظام‌مند در پایگاه‌های ScienceDirect، PubMed، IEEE Xplore، SpringerLink تا مرداد ۱۴۰۵ صورت گرفت. از میان ۱۹۹۰ رکورد اولیه، پس از حذف مقالات تکراری و غربالگری در سه مرحله (عنوان، عنوان و چکیده، متن کامل) در نهایت ۱۵ مطالعه برای تحلیل نهایی انتخاب شدند. به‌منظور افزایش جامعیت مرور، از دو روش جستجوی پسرو از طریق بررسی منابع مقالات منتخب و جستجوی پیشرو از طریق بررسی مقالات استنادکننده استفاده شد. با اعمال مجدد معیارهای ورود و خروج ۴ مطالعه دیگر شناسایی و به مجموعه نهایی افزوده شد. در نهایت ۱۵ مطالعه وارد مرور نظام‌مند شدند.

یافته‌ها: یادگیری ماشین پارادایمی پیشرفته در تولید لاگ رخداد مصنوعی است و بسته به نوع روش توسعه داده شده و روند تکاملی از روش‌های یادگیری ماشین سنتی (مانند درخت تصمیم)، تا روش‌های پیشرفته‌تر مبتنی بر یادگیری عمیق و هوش مولد (مانند مدل‌های زبانی بزرگ، اتوانکودرها و...) و روش‌های ترکیبی را در برمی‌گیرد. تحلیل مطالعات نشان‌دهنده توجه فزاینده به دسته دوم یعنی روش‌های مبتنی بر یادگیری عمیق و هوش مولد است. هدف اصلی تولید لاگ‌های رخداد مصنوعی ارزیابی الگوریتم (پنج مقاله (۳۳٪))، حفظ محرمانگی (پنج مقاله (۳۳٪))، تولید داده منفی (دو مقاله (۱۳٪))، تولید داده چند بعدی (دو مقاله (۱۳٪)) و در نهایت تحلیل سناریو (یک مقاله (۷٪)) بوده شده است. در تمام مقالات معیارهای وفاداری (یعنی سنجش شباهت آماری داده‌های مصنوعی به داده‌های واقعی و/یا شباهت ساختاری مسیرهای فرایند مصنوعی به مسیرهای واقعی) برای ارزیابی روش استفاده شده است، درحالی‌که کاربردپذیری در فرایند کاوی (به معنی ارزیابی روش‌های کشف یا بررسی انطباق با داده‌های مصنوعی تولید شده) در هشت مقاله (۵۳٪) و معیارهای محرمانگی تنها در دو مقاله (۱۳٪) گزارش شده است. یک تحلیل زمانی قابل توجه این است که از سال ۲۰۲۲ ابزارها کاملاً به سمت پیاده سازی تحت پایتون حرکت کرده‌اند؛ روندی که تا قبل از سال ۲۰۱۷ اصلاً وجود نداشته است.

بحث و نتیجه‌گیری: روش‌های مبتنی بر یادگیری عمیق و هوش مولد از سال ۲۰۲۳ مورد توجه فزاینده پژوهشگران قرار گرفته‌اند، درحالی‌که روش‌های ترکیبی کمتر بررسی شده‌اند. رویکردهای سنتی مبتنی بر شبیه‌سازی با بهره‌گیری از قوانین، الزامات و محدودیت‌های صریح کنترل‌جریان، صحت ساختاری و قابلیت حسابرسی بالایی فراهم می‌کنند؛ بااین‌حال، مهندسی پیچیده مدل، اسکرپت‌نویسی دستی و فرضیات ساده‌شده درباره لاگ‌های بدون ساختار می‌تواند به توزیع‌های تأخیر غیرواقعی منجر شود. در مقابل، روش‌های داده‌محور انعطاف‌پذیری بیشتری در یادگیری الگوهای پیچیده رفتاری، زمانی و چندبعدی دارند. GANها در مدل‌سازی مسیرهای موازی و وابستگی‌های دوربرد توانمندند، اما با ناپایداری آموزش و فروپاشی حالت مواجه می‌شوند. AE/CVAEها آموزش پایدار و فضاهای نهفته تفسیرپذیر ارائه می‌دهند، اما ممکن است رفتارهای نادر را کمتر بازنمایی کنند. DDPMها وفاداری بالایی ایجاد می‌کنند، ولی هزینه محاسباتی زیادی دارند. مدل‌های خودرگرسیو کنترل دقیق و انسجام زمانی را فراهم می‌کنند، درحالی‌که LLMها با وجود انعطاف‌پذیری بالا، مستعد توهم و نقض محدودیت‌های فرایند هستند. همچنین، فقدان چارچوب ارزیابی یکپارچه و ارزیابی محرمانگی، مقایسه روش‌ها را دشوار کرده است. پژوهش‌های آینده باید بر ترکیب مدل‌های مولد با مدل‌های ساختاریافته مانند شبکه‌های پتری و ارزیابی مقایسه‌ای مکانیسم‌های تولید تمرکز کنند تا مشخص شود کدام مکانیسم‌ها لاگ‌های مناسب‌تری برای فرایند کاوی تولید می‌کنند.

کلید واژه‌ها: لاگ رخداد مصنوعی، فرایند کاوی، یادگیری ماشین، یادگیری عمیق، مقاله مروری

A survey on machine learning-driven synthetic event log generation

Jalal Rezaeenour*, Mansoureh Yari Eili

PhD in Industrial Engineering, Professor, Department of Industrial Engineering, University of Qom,
Qom, Iran

PhD of Information Technology Engineering, Post Doc Researcher, Department of Industrial Engineering,
University of Qom, Qom, Iran

Email addresses: J.rezaee@qom.ac.ir, M.yari@stu.qom.ac.ir

Introduction: In recent years, process mining has emerged as a growing field within data science, playing an important role in extracting knowledge from data recorded in information systems. The event log constitutes the core of every process mining project. However, the limited availability of real-world data, due to privacy constraints, data collection costs, and poor data quality, poses significant challenges to the development of process mining algorithms. In this context, generating synthetic data that preserve the statistical and structural characteristics of processes while maintaining no direct correspondence with real-world instances offers an effective approach to addressing these challenges. Accordingly, this study aims to provide a systematic review of synthetic event log generation methods based on machine learning algorithms.

Methodology: The present study followed the systematic review framework proposed by Xiao and Watson (2019), which comprises protocol development, formulation of research questions, systematic literature searching, study screening, data extraction, and synthesis of findings. Accordingly, three research questions were defined: (1) What methods, algorithms, and machine learning-based approaches have been developed for synthetic event log generation in process mining? (2) What are the main applications of synthetic event logs in process mining studies? and (3) What methods have been used to evaluate their quality, utility, and performance? A systematic search was conducted in ScienceDirect, PubMed, IEEE Xplore, and SpringerLink through August 2026. From 1,990 initial records, duplicates were removed, followed by screening at three stages: title, title and abstract, and full text. Ultimately, 15 studies were selected for final analysis. To enhance the comprehensiveness of the review, backward and forward citation searching were also performed. Backward searching examined the reference lists of selected studies, while forward searching identified studies citing the selected articles. After reapplying the predefined inclusion and exclusion criteria, four

additional studies were identified and added to the final set. Thus, a total of 15 studies were included in the systematic review.

Findings: ML represents an advanced paradigm for synthetic event log generation. Depending on the underlying development approach, its evolution spans from conventional machine learning techniques, such as decision trees, to more advanced deep learning and generative artificial intelligence (GenAI) approaches, including LLMs and autoencoders, as well as hybrid methods that integrate multiple paradigms. The reviewed studies indicate a growing emphasis on the latter category, particularly deep learning- and GenAI-based approaches. The primary objectives of synthetic event log generation were algorithm evaluation (five studies, 33%), privacy preservation (five studies, 33%), negative data generation (two studies, 13%), multidimensional data generation (two studies, 13%), and, finally, scenario analysis (one study, 7%). All reviewed studies employed fidelity criteria to evaluate the proposed methods, referring to the assessment of the statistical similarity between synthetic and real data or the structural similarity between synthetic traces and the real one. In contrast, process-mining utility, defined as the evaluation of process discovery or conformance-checking techniques using the generated synthetic event logs, was reported in eight studies (53%), whereas privacy-related metrics were incorporated in only two studies (13%). A notable temporal trend is the complete shift toward Python-based implementations from 2022 onward, whereas such implementations were entirely absent from the reviewed studies published before 2017.

Discussion and Conclusion: Deep learning- and generative AI-based approaches have received increasing attention from researchers since 2023, whereas hybrid approaches have received comparatively limited attention. Traditional simulation-based approaches incorporate explicit rules, requirements, and control-flow constraints, thereby providing high structural validity and auditability. However, their reliance on complex model engineering, manual scripting, and simplified assumptions about unstructured event logs may result in unrealistic delay distributions. In contrast, data-driven approaches offer greater flexibility in learning complex behavioral, temporal, and multidimensional patterns. GANs are capable of modeling complex parallel paths and long-range dependencies but may suffer from training instability and mode collapse. AE/CVAE architectures provide stable training and interpretable latent spaces, although they may underrepresent rare behaviors. DDPMs can achieve high fidelity but entail substantial computational costs. Autoregressive models enable precise control and temporal coherence,

whereas LLMs, despite their high flexibility, remain susceptible to hallucinations and violations of process constraints. Moreover, the lack of an integrated evaluation framework and systematic privacy assessment remains a major challenge for comparing different approaches. Future research should focus on combining generative models with structured process models, such as Petri nets, and on comparative evaluations of generation mechanisms to determine which mechanisms produce event logs that are more suitable for process mining.

Keywords: Synthetic Event Log, Process Mining, Machine Learning, Deep Learning, Survey