Improving the Quality of Business Process Event Logs Using Unsupervised Method

Mohsen Mohammadi
Assistant Professor; Computer Department; Esfarayen University of Technology; Esfarayen, Iran Email: Mohsen@esfarayen.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
Special Issue | Winter 2025 | pp. 21-40
Exploring the Relationship Between Data Quality and Business Process Management
https://doi.org/10.22034/jipm.2025.2018045.1469

Received: 16, Dec. 2023 **Accepted: 06, Feb. 2024**

Abstract: In the contemporary dynamic business environment, the dependability of process mining algorithms is intricately tied to the quality of event logs, often marred by data challenges stemming from human involvement in business processes. This study introduces a novel approach that amalgamates insights from prior works with unsupervised techniques, specifically Principal Component Analysis (PCA), to elevate the precision and reliability of event log representations. Executed through Python and the pm4py library, the methodology is applied to real event logs. The adoption of Petri nets for process representation aligns with systematic approaches advocated by earlier studies, enhancing transparency and interpretability. Results demonstrate the method's efficacy through enhanced metrics such as Fitness, Precision, and F-Measure, accompanied by visualizations elucidating the optimal number of principal components. This study offers a comprehensive and practical solution, bridging gaps in existing methodologies, and its integration of multiple strategies, particularly PCA, showcases versatility in optimizing process mining analyses. The consistent improvements observed underscore the method's potential across diverse business contexts, making it accessible and pertinent for practitioners engaged in real-world business processes. Overall, this research contributes an innovative approach to improve event log quality, thereby advancing the field of process mining with practical implications for organizational decision-making and process optimization.

Keywords: Process Mining, Quality Metrics, Business Process Model, Event Log

* Corresponding Author

1. Introduction

In the current dynamic business landscape, where uncertainties in business process management often result in frequent shifts in requirements, making the complexity of business processes increasingly evident.

In this context, process mining algorithms utilize event logs to reveal hidden insights within various business processes. These logs consist of data that detail the execution of processes recorded by the information systems that support the respective business processes (Dumas et al., 2018; van der Aalst, 2016). Despite the significant potential of process mining to improve organizational understanding and optimize processes, the reliability of its outcomes is closely linked to the quality of the event log (Bose et al., 2013; van der Aalst et al., 2016).

Real-world event logs frequently encounter various data quality challenges, such as missing events, inaccurate timestamps, and erroneous resource information (Suriadi et al., 2017). Many of these challenges stem from human involvement in business processes, which introduces risks such as delayed, erroneous, and incomplete data recording.

Deploying an event log that is fraught with data quality issues without careful consideration may yield counterintuitive or misleading process mining results, potentially leading to suboptimal or even detrimental managerial decisions (Mohammadi, 2017; Andrews et al., 2018; Martin et al., 2022).

Therefore, effectively addressing issues such as noise, outliers, and imperfections in event logs is essential for deriving valuable insights, facilitating informed decision-making, and navigating the complexities of today's dynamic business landscape.

2. Literture review

The succeeding subsections present the contextual information, segmented into two components. The first part encompasses the definitions of process mining and event logs, along with the associated quality metrics, which will be the central focus of this study. The second part offers a glimpse into pertinent previous studies.

2-1. Background

These logs are typically generated by process-aware information systems, such as customer relationship management (CRM) systems, enterprise resource planning (ERP) systems, and business process management systems (BPMS). In the context of Business Process Management (BPM), process mining is a valuable technique, as illustrated in Figure 1. It encompasses three key types: process improvement, process discovery, and conformance checking. Process discovery involves creating process models from the data stored in event logs, enabling organizations to understand how their processes actually function. Conformance checking, on the other hand, focuses on comparing event logs with the original process model to identify any deviations or discrepancies.

Finally, process improvement seeks to enhance existing process models by incorporating insights from event logs, thereby increasing the overall efficiency and effectiveness of business processes (van der Aalst, 2016).

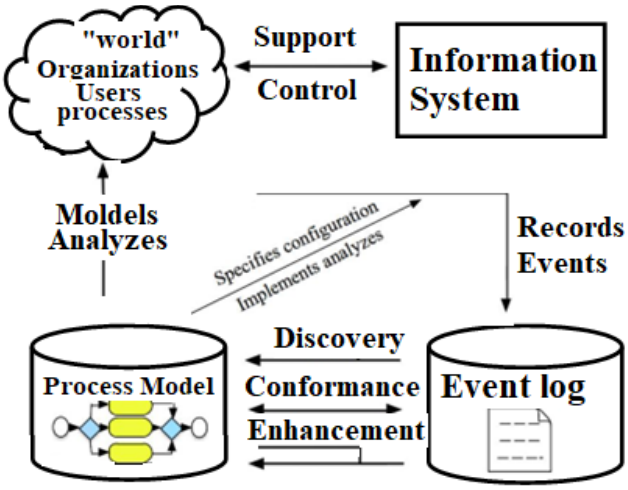

Figure 1. Process mining in BPM (van der Aalst 2016)

The initial stage of process mining involves obtaining an event log, which may originate from an actual information system log or be created from historical data stored in a database. Regardless of its origin, this event log must adhere to a predetermined structure. For subsequent analysis using process mining methodologies, events of interest—representing user actions within process instances—must be meticulously documented. Once a significant volume of

these events is accumulated, they collectively constitute an event log. To facilitate effective process mining, each event within this log should include essential details, such as a case ID that identifies the process instance, a task name that specifies the activity, a user name that identifies the participant, and a timestamp that indicates when the task was completed. An ideal event log for process mining is customized for a specific process and is structured as a collection of cases or traces in a multi-record case format, as illustrated in Figure 2. The essential information contained in each event, including the case ID, task name, and user name, ensures that the log is suitable for thorough process analysis (Ferreira, 2017; Mohammadi, 2019).

CaseID	Action	User	Task	lifecycle:transition	time:timestamp
0	Created	User_1	A_Create Application	complete	2016-01-01 09:51:15.304000+00:00
1	statechange	User_1	A_Submitted	complete	2016-01-01 09:51:15.352000+00:00
2	Created	User_1	W_Handle leads	schedule	2016-01-01 09:51:15.774000+00:00
3	Deleted	User_1	W_Handle leads	withdraw	2016-01-01 09:52:36.392000+00:00
4	Created	User_1	W_Complete application	schedule	2016-01-01 09:52:36.403000+00:00

Figure 2. sample of an event log.

In an event log, each case represents the sequence of events executed during a single iteration of a process instance. A variant is defined as a unique sequence of events from the start to the end of a process instance. Each case or trace corresponds to a specific variant, and a variant may consist of one or more cases or traces (Suriadi, 2017). The process of acquiring an event log, which is fundamental to process mining, is explained here, emphasizing the importance of maintaining a structured format. These event logs, organized into cases or traces, form the basis for comprehensive process analysis. To assess the effectiveness of a process model in accurately capturing observed behavior, four distinct dimensions of quality are considered: simplicity, replay fitness, precision, and generalization.

Simplicity assesses a model's comprehensibility for humans and is independent

of the observed behavior. Since different process models can represent the same behavior in various ways, it is advisable to prioritize the simplest model (Van der Aalst et al., 2012; Van der Aalst, 2016). Replay fitness, another quality dimension, quantifies the proportion of behavior documented in the event log that can be accurately reproduced by the process model. In contrast, precision measures the extent of behavior permitted by the process model that is not observed in the event log. Both replay fitness and precision assess the alignment between the event log and the process model. It is important to note that the event log captures only a fraction of the potential behaviors permitted by the system.

Consequently, the generalization dimension evaluates whether the process model avoids excessive customization to the observed event log behavior and accurately represents the broader system. Essentially, generalization also gauges the model's ability to describe behaviors that have not yet been observed within the system (Van Der Aalst, 2013; Buijs et al., 2014).

In the context of process mining, quality metrics for event logs prioritize replay fitness and precision over simplicity and generalization because of their direct influence on the accuracy and reliability of the discovered process models.

Replay fitness is crucial as it assesses how well the model aligns with actual event data during replay, ensuring fidelity to observed behavior. Precision, conversely, assesses the accuracy of the model's predictions concerning subsequent activities in a process, thereby impacting its practical utility. While simplicity and generalization are important, prioritizing replay fitness and precision is essential for ensuring that the process model faithfully represents real-world processes and makes precise predictions. This approach provides a more robust foundation for process optimization and decision-making in practical applications (van der Aalst, 2018; Sani, 2020).

Related works

The quality of event logs in the context of process mining has been the subject of extensive investigation, with researchers striving to develop systematic approaches for cleaning and evaluating imperfections in event logs.

In the work of Suriadi et al. (2017), a significant contribution was made toward identifying imperfection patterns in event logs. This study represents a

step forward in the development of systematic techniques for cleaning event logs, offering insights into common imperfections that can affect the accuracy of subsequent process mining analyses. Janssenswillen et al. (2017) conduct a comparative study of existing quality measures for process discovery, enhancing the understanding of various metrics and their applicability in evaluating the quality of discovered process models.

This comparative analysis assists in identifying the most effective strategies for ensuring the reliability of process mining outcomes. Furthermore, the comprehensive review by Koschmider et al. (2021) highlights the challenges posed by noise and outliers in event logs, offering valuable insights for future research directions. While not proposing a specific method, their work enhances our understanding of the challenges related to event log quality. Anomalies and outliers in event logs can significantly distort subsequent analyses and hinder the development of reliable process models. Nolle et al. (2016) demonstrated that denoising auto encoders are effective in managing noisy business process event logs through unsupervised learning, offering flexibility and adaptability to various datasets. However, challenges such as the interpretability of the learned representations and the potential for overfitting are significant concerns. Additionally, Nguyen et al. (2019) focused on using auto encoders to improve event log quality, providing a more structured representation of the underlying processes. While auto encoders are proficient at capturing complex patterns, their performance can be influenced by the choice of hyper parameters, and their effectiveness may vary across different datasets. (Fani Sani et al., 2018) explore sequence mining for outlier detection, emphasizing the importance of considering the sequential nature of process data. This approach may excel in scenarios where temporal dependencies are crucial. However, its effectiveness may be influenced by the complexity and variability of process sequences.

In their 2021 study, Ko and Comuzzi present a statistical leverage-based approach that provides an alternative perspective for anomaly detection, grounded in statistical principles. However, the effectiveness of this method may depend on the suitability of the selected statistical measures and could be sensitive to the underlying data distribution.

Marin-Castro and Tello-Leal (2021) conduct a comprehensive review of

event log preprocessing methods within the field of process mining. They explore various techniques designed to refine event logs, thereby enhancing their quality and reliability—factors that are crucial for effective process analysis and improvement. The review identifies common challenges, such as noise and outliers in event logs, and examines strategies to mitigate these issues. By synthesizing existing research, the review emphasizes the significance of robust event log preprocessing for achieving accurate outcomes in process mining and highlights potential avenues for future investigation in this area.

Goel et al. (2022) advocate for standardized data quality annotations within the process mining workflow to enhance the reliability of analytical outcomes. They propose systematically integrating data quality metrics into process mining, emphasizing clarity and consistency in the annotation process. Through a literature review and empirical evidence, this study underscores the benefits of quality-informed process mining in enhancing decision-making and optimizing processes. By promoting standardized data quality annotations, the authors aim to align process mining practices with industry standards and enhance data quality management.

Martin et al. (2022) introduce Daqapo, a flexible framework designed to assess the quality of event logs in process mining endeavors. Acknowledging the critical importance of high-quality event logs, they propose Daqapo as a tool for evaluating various aspects of event log quality in a detailed and nuanced manner. The framework enables the customization of quality metrics to align with specific process mining objectives, thereby facilitating informed decision-making. By systematically analyzing the characteristics of event logs, Daqapo assists in identifying data quality issues and improving the reliability of process mining initiatives. Bayomie et al. (2023) present a novel approach to event-case correlation in process mining by utilizing probabilistic optimization techniques. Their study focuses on establishing strong connections between individual events and the corresponding cases or processes they represent. By integrating probabilistic optimization methods, this research aims to enhance the accuracy and efficiency of event-case correlation, resulting in more insightful process analyses and improved decision-making. The findings contribute to the advancement of process mining methodologies by addressing key challenges in event-case correlation and enriching the field of process mining research and practice. The research gap

focuses on addressing noise, outliers, and anomalies in event logs to enable robust process mining analyses. While existing studies have contributed to identifying patterns of imperfection and comparing quality measures, there remains a need to integrate unsupervised methods, such as Principal Component Analysis (PCA), to further enhance the quality of event logs.

Despite previous explorations of techniques such as denoising autoencoders and sequence mining, challenges such as interpretability and dataset variability persist. Incorporating unsupervised methods like Principal Component Analysis (PCA) is essential, as it facilitates dimensionality reduction, simplifies complex datasets, and enhances computational efficiency. PCA's unsupervised nature enables the identification of hidden patterns without the need for labeled data, aligning with the objective of ensuring accurate representations of the underlying processes in event logs.

By integrating unsupervised techniques with insights from previous studies, a comprehensive approach can be developed to address noise, outliers, and anomalies in event logs, thereby enhancing process mining analyses.

3. Method

As mentioned earlier, process mining is a methodological approach used to extract insights and construct process models from event logs generated by information systems. It involves analyzing event data to understand how processes function in real-world scenarios. This section provides a comprehensive overview of the steps that constitute the proposed method, as illustrated in Figure 3. The proposed methodology relies on key tools, particularly Python and the pm4py library. Python is a versatile and widely used programming language that offers a robust ecosystem of libraries for data manipulation, analysis, and visualization. Pm4py, a specialized Python library, includes a diverse range of process mining algorithms, providing a comprehensive toolkit for managing event logs, extracting valuable insights, and constructing process models (Berti et al., 2019). The selection of the Loan Application Process from a Dutch financial institution (BPI Challenge 2017) and the Purchase Order Handling Process (BPI Challenge 2019) from the IEEE Task Force on Process Mining (https://community.data.4tu.nl) was based on their relevance and complexity within the financial and procurement sectors. The loan

application process involves several stages, including application submission, verification, approval, and disbursement, which reflect a typical workflow in financial institutions. Similarly, the Purchase Order Handling Process includes steps such as order creation, approval, fulfillment, and invoicing, which are essential operations in procurement activities. These processes provide extensive datasets featuring a variety of event sequences, enabling thorough analysis and insights into process dynamics. The decision to utilize Petri nets as the representation for discovered process models is based on their ability to provide a concise and transparent visualization of process behavior. Petri nets offer a standardized framework that is widely accepted in the Business Process Management (BPM) domain, facilitating clear representation and analysis of process flows and dependencies. Their ability to capture all traces from the initial marking to the final marking makes them particularly well-suited for modeling the complexities of business processes in process mining analyses (Boltenhagen et al., 2019).

In the described process, Principal Component Analysis (PCA) is employed to reduce the dimensionality of the event log data (Kurita, 2019) capturing the most significant variations while preserving its essential characteristics. PCA aids in determining the optimal number of components, thereby streamlining the data for further analysis. Following Principal Component Analysis (PCA), the Isolation Forest algorithm is utilized to detect and remove outliers from the event log. Isolation Forest is a machine learning technique that identifies anomalies by isolating them into smaller partitions. By integrating PCA with Isolation Forest, this method effectively preprocesses the event log, ensuring it contains high-quality data that is suitable for subsequent analysis. Finally, quality metrics are calculated to evaluate the cleaned event log, offering insights into its reliability and appropriateness for process mining tasks. In the context of the study, an accepting Petri net refers to a specific type of Petri net that encapsulates the complete behavior of a process from its initial state to its final state. This encompasses all possible sequences of events or activities that may occur within the process. The decision to adopt Petri nets in this study is essential, as they provide a comprehensive representation of process behavior, ensuring that no potential sequence of events is overlooked during the modeling process. Furthermore, Petri nets offer a clear and unambiguous method for visualizing and analyzing

process flow, making them particularly well-suited for process mining applications. The widespread adoption of Business Process Management (BPM) and Workflow Management (WFM) systems underscores their relevance and effectiveness in accurately and comprehensively representing process models within the context of this study (Van der Aalst, 2016).

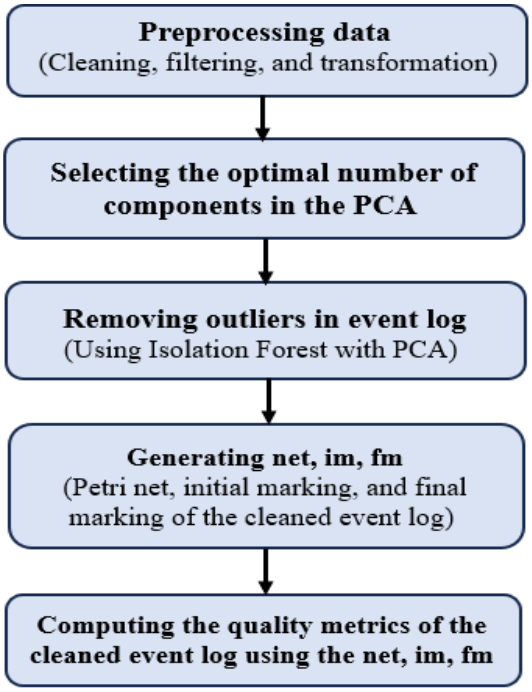

Figure 3. the proposed method

Referring to Figure 3, the initial step of the process involves clearing and filtering the event log associated with the business process. Following this, the data is converted into a dataframe in preparation for the subsequent phase, which entails determining the optimal number of components in the PCA corresponding to the event log. In the subsequent stage, utilizing the optimal number of components, Isolation Forest in conjunction with PCA is employed to eliminate outliers from the event log. Finally, quality metrics are computed for the refined event log.

Regarding the third step outlined in the proposed methodology, PCA serves as a fundamental tool in contemporary data analysis, spanning various scientific domains. Its primary objective is to identify a meaningful basis for re-expressing

datasets, revealing underlying structures while minimizing noise. PCA finds utility in tasks such as dimensionality reduction, data compression, feature extraction, and data visualization, showcasing its adaptability across diverse analytical contexts. Furthermore, Isolation Forest functions as an unsupervised anomaly detection method, capitalizing on the notion that anomalies are likely to be isolated with fewer splits in a random dataset partition. By constructing a binary tree structure and isolating instances along shorter paths, anomalies are efficiently identified owing to their distinctiveness. This unsupervised methodology excels in outlier detection without necessitating labeled data, making it particularly advantageous in scenarios where anomalies may lack clear definition (Post et al 2021).

It should be mentioned that in the fourth step of the proposed method, Petri nets play a crucial role due to their ability to encapsulate the behavior of an accepting Petri net, covering all traces from the initial marking to the final marking. Furthermore, Petri nets are the chosen means for representing the discovered process models, owing to their brevity and straightforward, unambiguous semantics. In the domain of process mining, Petri nets stand out as the most widely adopted representation, forming the foundation for process models in both BPM (Business Process Management) and WFM (Workflow Management) systems (Van der Aalst 2016).

In the context of process mining, fitness and precision are metrics used to evaluate the quality of event logs and process models. Fitness measures how well a process model reproduces the observed behavior recorded in the event log. It quantifies the degree to which the model aligns with the actual process executions. Precision, on the other hand, assesses the accuracy of predictions made by the process model regarding the next activities in a process. It measures the extent to which the model's predictions match the observed behavior in the event log. There is often a trade-off between fitness and precision; while optimizing one metric may improve its value, it can negatively impact the other. The F-Measure metric is commonly used to strike a balance between fitness and precision by combining both metrics into a single score, offering a comprehensive assessment of the process model's performance. The F-Measures metric is utilized, combining both metrics through the formula (van der Aalst 2018; Fani Sani 2020):

$$\text{F-Measures} = (2 \times \text{Precision} \times \text{Fitness}) / (\text{Precision} + \text{Fitness})$$

4. Findings and Discussion

Improving the quality of business process event logs is paramount for accurate process mining analyses and informed decision-making. High-quality event logs offer a reliable depiction of process executions, enabling organizations to identify inefficiencies, optimize workflows, and enhance operational efficiency. Moreover, accurate event logs aid in demonstrating compliance with regulatory requirements, mitigating risks, and ensuring accountability. By prioritizing strategies to enhance event log quality, organizations can gain valuable insights, bolster decision-making capabilities, and foster sustainable growth in today's competitive business landscape.

The proposed method, as delineated in Figure 3, has proven to be highly effective in augmenting the quality of event logs, with a specific focus on two distinct business processes—the Loan Application Process of a Dutch financial institute (BPI Challenge, 2017) and the Purchase Order Handling Process (BPI Challenge ,2019). Key evaluation metrics, including Fitness, Precision, and F-Measure, serve as vital indicators to gauge the impact of the proposed method on the quality of these event logs.

Figures 4 and 5 provide valuable insights into how Principal Component Analysis (PCA) contributes to the dimensionality reduction process within the proposed methodology. The plots illustrate the relationship between the number of principal components and the cumulative explained variance, which indicates the extent to which the components capture variability in the event logs. As more components are included, the cumulative explained variance increases, reflecting the proportion of data variability captured by the PCA model. However, there comes a point at which adding more components leads to diminishing returns in capturing additional variance. Therefore, identifying the optimal number of components is crucial for balancing the retention of essential information with computational efficiency.

In the BPI 2017 event log, five components are considered optimal, whereas the BPI 2019 log achieves the best balance with fifteen components. This understanding guides subsequent analysis steps, such as dimensionality reduction and outlier removal, ensuring that the event logs are of high quality and reliability for process mining applications.

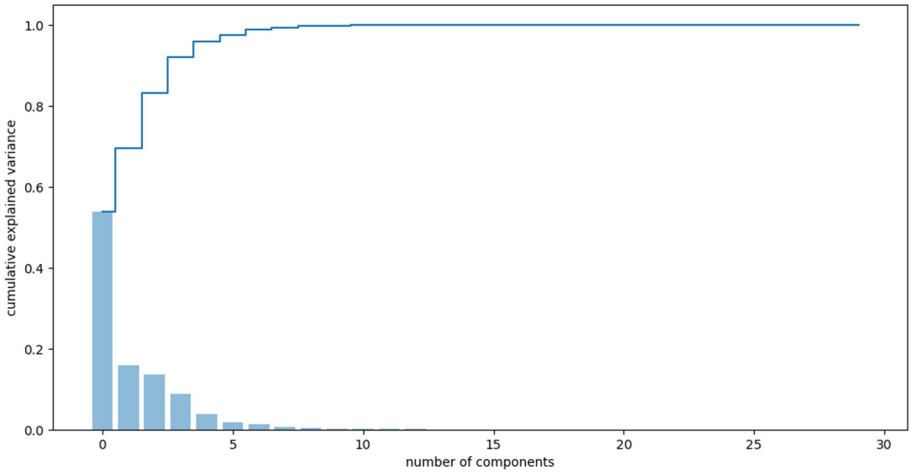

Figure 4. number of principal components in BPI 2017

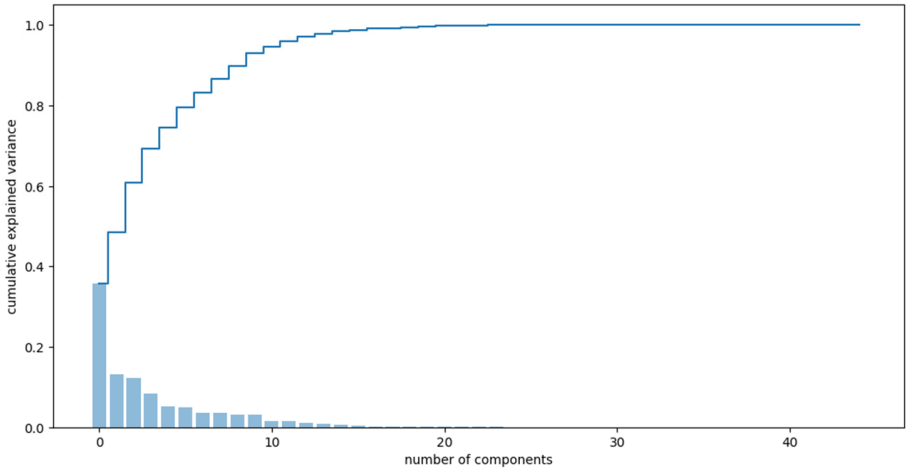

Figure 5. Number of principal components in BPI 2019

Table 1. The quality metrics results for BPI 2017 and BPI 2019

Event log	Number of Cases	Fitness	Precision	F-Measure
BPI 2017	31509	0.963	0.814	0.882
BPI 2017- Cleaned	30166	0.965	0.940	0.952
BPI 2019	251734	0.915	0.872	0.893
BPI 2019- Cleaned	239161	0.942	0.941	0.941

Table 1 serves as a crucial reference point for evaluating the effectiveness of the proposed method on the quality metrics of the BPI 2017 and BPI 2019 event logs. These metrics play a pivotal role in assessing the accuracy and reliability of process mining outcomes derived from the event logs. Fitness, which measures the alignment between discovered process models and real executions, exhibited a marginal improvement in the BPI 2017 log, rising from 0.963 to 0.965 post-application of the method. This signifies a closer correspondence between the models and actual processes. Notably, Precision, reflecting the model's ability to distinguish true positives while minimizing false positives, experienced a significant enhancement in the BPI 2017 log, escalating from 0.814 to 0.940. Consequently, the F-Measure, a composite metric balancing precision and recall, demonstrated a substantial optimization, climbing from 0.882 to 0.952. Similarly, for the BPI 2019 log, improvements were evident across all metrics. Fitness surged from 0.915 to 0.942, while Precision exhibited a minor uplift from 0.872 to 0.941, resulting in an F-Measure of 0.941. These enhancements underscore the method's efficacy in refining event log quality, bolstering the reliability and accuracy of process mining analyses across various business domains.

The consistent improvements observed in both the BPI 2017 and BPI 2019 event logs indicate a robust and adaptable approach within the proposed methodology, highlighting its versatility across various business processes.

By utilizing Petri nets to represent the discovered process models, this methodology ensures both clarity and compliance with industry standards in Business Process Management (BPM) and Workflow Management (WFM) systems. The integration of Principal Component Analysis (PCA) and Isolation Forest techniques is a crucial element of the method, effectively addressing outliers and significantly improving the overall quality of event logs. This highlights the significance of utilizing unsupervised methods to enhance event log data, establishing a robust foundation for more precise process mining analyses. Moreover, the comprehensive nature of this approach, which seamlessly integrates Python functionalities with specialized libraries, provides a holistic strategy for data refinement.

This integration facilitates a more comprehensive and systematic analysis of event logs, enhancing the reliability and accuracy of process mining results.

Notably, this method marks a significant advancement over previous studies, which frequently depended on isolated techniques such as denoising autoencoders or statistical leverage-based anomaly detection. By leveraging Principal Component Analysis (PCA) for dimensionality reduction, this methodology enhances computational efficiency by customizing the reduction process to the unique characteristics of each event log.

Additionally, the use of Petri nets for representing process models enhances transparency and interpretability, aligning effectively with established systematic approaches in the field. Overall, the observed improvements highlight the potential of the proposed method to transform business process analysis and decision-making. By offering clearer insights into process dynamics and outliers, this methodology enables organizations to make more informed and strategic decisions, ultimately enhancing efficiency and fostering innovation across various industry sectors.

The proposed method signifies a substantial advancement in process mining, especially in tackling the challenges highlighted by previous research. Unlike some research that is primarily theoretical or technique-focused, our method effectively bridges the gap between theory and practice by providing tangible solutions to real-world problems.

Moreover, while Koschmider et al. (2021) identified challenges related to noise and outliers without proposing specific solutions, our reliance on established tools such as Python and the pm4py library reflects a comprehensive strategy that makes our method accessible to practitioners. In contrast to approaches like denoising autoencoders (Nolle et al., 2016) or sequence mining (Fani Sani et al., 2018), which may encounter interpretability issues or exhibit limited effectiveness across various datasets, our method integrates unsupervised techniques such as Principal Component Analysis (PCA) to enhance the quality of event logs, thereby improving the effectiveness of process mining analyses. Additionally, our approach builds upon the insights provided by Marin-Castro and Tello-Leal (2021) and Martin et al. (2022) by offering a practical framework for event log preprocessing and quality assessment. By integrating insights from previous studies with practical applications, our method addresses a significant gap in the existing literature and provides a more robust approach to managing noise, outliers, and anomalies in event logs, ultimately improving process mining analyses.

5. Conclusion

In conclusion, this study presents a groundbreaking methodology that combines Principal Component Analysis (PCA) and Isolation Forest to improve the quality of event logs in process mining applications. PCA, a dimensionality reduction technique, facilitates the identification of the most significant variations in the event log data while preserving essential characteristics. By reducing the dimensionality of the data, Principal Component Analysis (PCA) streamlines subsequent analyses and enhances computational efficiency.

Isolation Forest, an unsupervised anomaly detection algorithm, effectively identifies and removes outliers from the event log, thereby improving its quality and reliability. The integration of Principal Component Analysis (PCA) and Isolation Forest provides several advantages. Firstly, PCA simplifies complex datasets by capturing essential variations, making them more manageable for analysis. This reduction in dimensionality not only accelerates computational processes but also improves the interpretability of the data.

Secondly, the Isolation Forest algorithm excels in outlier detection without the need for labeled data, making it particularly useful in scenarios where anomalies may not have clear definitions. By identifying and removing outliers, the Isolation Forest ensures that the event log is free from irregularities, leading to more accurate process mining results. The quantitative results from the study demonstrate the effectiveness of the proposed methodology in enhancing the quality of event logs. For instance, in the Loan Application Process (BPI Challenge 2017) and the Purchase Order Handling Process (BPI Challenge 2019), significant improvements were observed in key metrics such as Fitness, Precision, and F-Measure after applying the methodology. In the BPI 2017 event log, Fitness increased from 0.963 to 0.965, Precision improved from 0.814 to 0.940, and the F-Measure rose from 0.882 to 0.952 following the application of the method. Similarly, in the BPI 2019 event log, Fitness increased from 0.915 to 0.942, Precision improved from 0.872 to 0.941, and the F-Measure reached 0.941. These quantitative results underscore the effectiveness of the methodology in enhancing event log quality and its potential for practical application in real-world business processes. In summary, the integration of Principal Component Analysis (PCA) and Isolation Forest offers a robust method for preprocessing

event logs in process mining applications. This methodology not only enhances key performance metrics but also demonstrates its applicability across various business processes, underscoring its practical significance.

Limitation of the Study:

Despite the significant strides made in enhancing event log quality through our methodology, it is essential to acknowledge its limitations. In scenarios characterized by high complexity and variability in process sequences, our approach may face challenges. These challenges arise from the complexities involved in accurately interpreting and processing diverse process patterns. Furthermore, although our methodology proves effective across various datasets, there are instances where the interpretability of the denoising process and the sensitivity of certain hyperparameters present significant concerns. Addressing these challenges is essential for improving the robustness and applicability of the methodology in real-world scenarios.

Future Research Directions:

Moving forward, several avenues for future research warrant exploration to refine and enhance our methodology.

One critical aspect involves further refining the methodology to more effectively manage high-complexity process sequences. This involves creating strategies to effectively interpret and analyze complex process patterns, thereby improving the methodology's adaptability to various scenarios. Exploring the integration of advanced machine learning techniques for adaptive hyperparameter tuning represents a promising direction. By leveraging machine learning algorithms, we can dynamically adjust hyperparameters to optimize the methodology's performance across diverse datasets and scenarios. Furthermore, expanding the application of our approach to a broader array of industry-specific processes will yield valuable insights into its generalizability and effectiveness across various domains.

Finally, it is essential to investigate the scalability of the method to handle larger datasets and real-time event log processing. This will allow the methodology to adapt to the evolving needs of process mining applications, enabling more comprehensive and sophisticated analyses.

References

Andrews, R., Suriadi, S., Ouyang, C., & Poppe, E. (2018). Towards event log querying for data quality: Let's start with detecting log imperfections. In On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part I (pp. 116-134). Springer International Publishing. DOI: 10.1007/978-3-030-02610-3_7

Bayomie, D., Di Ciccio, C., & Mendling, J. (2023). Event-case correlation for process mining using probabilistic optimization. Information Systems, 114, 102167. DOI: 10.1016/j.is.2023.102167

Berti, A., Van Zelst, S. J., & van der Aalst, W. (2019). Process mining for python (PM4Py): bridging the gap between process-and data science. arXiv preprint arXiv:1905.06169. DOI: 10.48550/arXiv.1905.06169

Boltenhagen, M., Chatain, T., & Carmona, J. (2019). Generalized alignment-based trace clustering of process behavior. In Application and Theory of Petri Nets and Concurrency: 40th International Conference, PETRI NETS 2019, Aachen, Germany, June 23–28, 2019, Proceedings 40 (pp. 237-257). Springer International Publishing. DOI: 10.1007/978-3-030-21571-2_

Bose, R. J. C., Mans, R. S., & Van Der Aalst, W. M. (2013, April). Wanna improve process mining results?. In 2013 IEEE symposium on computational intelligence and data mining (CIDM) (pp. 127-134). IEEE. DOI: 10.1109/CIDM.2013.6597227

Buijs, J. C., van Dongen, B. F., & van der Aalst, W. M. (2014). Quality dimensions in process discovery: The importance of fitness, precision, generalization and simplicity. International Journal of Cooperative Information Systems, 23(01), 1440001. DOI: 10.1142/S0218843014400012

Dumas, M., Rosa, L. M., Mendling, J., & Reijers, A. H. (2018). Fundamentals of business process management. Springer-Verlag. DOI: 10.1007/978-3-662-56509-4

Ferreira, D. R. (2017). A primer on process mining: Practical skills with python and graphviz. Cham: Springer International Publishing. DOI: 10.1007/978-3-030-41819-9

Goel, K., Leemans, S. J., Martin, N., & Wynn, M. T. (2022). Quality-informed process mining: A case for standardised data quality annotations. ACM Transactions on Knowledge Discovery from Data (TKDD), 16(5), 1-47. DOI: 10.1145/3511707

Janssenswillen, G., Donders, N., Jouck, T., & Depaire, B. (2017). A comparative study of existing quality measures for process discovery. Information Systems, 71, 1-15. DOI: 10.1016/j.is.2017.06.002

Ko, J., & Comuzzi, M. (2021). Detecting anomalies in business process event logs using statistical leverage. Information Sciences, 549, 53-67. DOI: 10.1016/j.ins.2020.11.017

Koschmider, A., Kaczmarek, K., Krause, M., & van Zelst, S. J. (2021, September). Demystifying noise and outliers in event logs: Review and future directions. In International Conference on Business Process Management (pp. 123-135). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-94343-1_10

Kurita, T. (2021). Principal component analysis (PCA). In Computer vision: a reference guide (pp. 1013-1016). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-63416-2_649

Martin, N., Van Houdt, G., & Janssenswillen, G. (2022). DaQAPO: supporting flexible and fine-grained event log quality assessment. Expert Systems with Applications, 191, 116274. DOI: 10.1016/j.eswa.2021.116274

Marin-Castro, H. M., & Tello-Leal, E. (2021). Event log preprocessing for process mining: a review. Applied Sciences, 11(22), 10556. DOI: 10.3390/app112210556

Mohammadi, M. (2017). A Review of influencing factors on the quality of business process models. Journal of Economic & Management Perspectives, 11(3), 1833-1840.

Mohammadi, M. (2019, September). Discovering business process map of frequent running case in event log. In 2019 international conference on information technologies (InfoTech) (pp. 1-4). IEEE. DOI: 10.1109/InfoTech.2019.8860877

Nolle, T., Seeliger, A., & Mühlhäuser, M. (2016). Unsupervised anomaly detection in noisy business process event logs using denoising autoencoders. In Discovery Science: 19th International Conference, DS 2016, Bari, Italy, October 19–21, 2016, Proceedings 19 (pp. 442-456). Springer International Publishing. DOI: 10.1007/978-3-319-46307-0_28

Nguyen, H. T. C., Lee, S., Kim, J., Ko, J., & Comuzzi, M. (2019). Autoencoders for improving quality of process event logs. Expert Systems with Applications, 131, 132-147. DOI: 10.1016/j.eswa.2019.04.052

Post, R., Beerepoot, I., Lu, X., Kas, S., Wiewel, S., Koopman, A., & Reijers, H. (2021, October). Active anomaly detection for key item selection in process auditing. In International Conference on Process Mining (pp. 167-179). Cham: Springer International Publishing. DOI: 10.1007/978-3-030-98581-3_13

Fani Sani, M., van Zelst, S. J., & van der Aalst, W. M. (2018). Applying sequence mining for outlier detection in process mining. In On the Move to Meaningful Internet Systems. OTM 2018 Conferences: Confederated International Conferences: CoopIS, C&TC, and ODBASE 2018, Valletta, Malta, October 22-26, 2018, Proceedings, Part II (pp. 98-116). Springer International Publishing. DOI: 10.1007/978-3-030-02671-4_6

Sani, M. F. (2020, June). Preprocessing event data in process mining. In CAiSE (Doctoral Consortium) (pp. 1-10).

Suriadi, S., Andrews, R., ter Hofstede, A. H., & Wynn, M. T. (2017). Event log imperfection patterns for process mining: Towards a systematic approach to cleaning event logs. Information systems, 64, 132-150. DOI: 10.1016/j.is.2016.07.011

Van der Aalst, W., Adriansyah, A., & Van Dongen, B. (2012). Replaying history on process models for conformance checking and performance analysis. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(2), 182-192. DOI: 10.1002/widm.1045

Van Der Aalst, W., & van der Aalst, W. (2016). Data science in action (pp. 3-23). Springer Berlin Heidelberg.DOI: 10.1007/978-3-662-49851-4_1

van der Aalst, W. M. (2013, May). Mediating between modeled and observed behavior: The quest for the "right" process: keynote. In IEEE 7th International Conference on Research Challenges in Information Science (RCIS) (pp. 1-12). IEEE. DOI: 10.1109/RCIS.2013.6577675

Sani, M. F., van Zelst, S. J., & Van Der Aalst, W. M. (2018). Improving process discovery results by filtering outliers using conditional behavioural probabilities. In Business Process Management Workshops: BPM 2017 International Workshops, Barcelona, Spain, September 10-11, 2017, Revised Papers 15 (pp. 216-229). Springer International Publishing. DOI: 10.1007/978-3-319-74030-0_16

Van Der Aalst, W., Adriansyah, A., De Medeiros, A. K. A., Arcieri, F., Baier, T., Blickle, T., ... & Wynn, M. (2012). Process mining manifesto. In Business Process Management Workshops: BPM 2011 International Workshops, Clermont-Ferrand, France, August 29, 2011, Revised Selected Papers, Part I 9 (pp. 169-194). Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-28108-2_19

Mohsen Mohammadi

Graduated from the National University of Malaysia (UKM) with a PhD in Information Technology (Industrial Computing) in 2014. He is currently an assistant professor at Esfarayen University of Technology, Iran.

Process mining, Business processes analysis and modeling, and Information systems design are his favorite research fields.