IJIRSET
International Journal of Innovative Research in
SCIENCE | ENGINEERING | TECHNOLOGY

e-ISSN: 2319-8753 | p-ISSN: 2347-6710

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

Volume 12, Issue 11, November 2023

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.423

9940 572 462 | 6381 907 438 | ijirset@gmail.com | www.ijirset.com

International Journal of Innovative Research in Science, Engineering and Technology (IJIRSET)

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

Adaptive Feature Extraction Approaches for Managing Concept Drifts in Process Mining

Puneetha B H, Puneeth S P

Assistant Professor , Department of Computer Science & Business systems, Bapuji Institute of Engineering & Technology, Davangere, Karnataka, India

Assistant Professor , Department of Information Science & Engineering, Bapuji Institute of Engineering & Technology, Davangere, Karnataka, India

ABSTRACT: Process mining acts as a vital link between data mining and business process modeling, with the primary objective of uncovering various process models based on historical or event log data. Concept drift, in this context, signifies the scenario where the process undergoes changes during analysis. Current strategies and analysis techniques in process mining often fall short when confronted with processes experiencing such drift. Conventional process mining methods tend to treat processes as static entities, assuming that they remain consistent throughout their execution. This paper focuses on presenting a novel approach to detect and localize concept drift from a control-flow perspective using a feature extraction mechanism. The proposed method introduces a set of distinct features designed to capture the dynamic relationships among activities in a process.

KEYWORDS: Concept Drift, Process Mining, Process, Drift, Feature Extraction

I. INTRODUCTION

Process mining is a powerful process management technique that facilitates the analysis of business processes using event logs. It involves extracting valuable insights from event logs generated by information systems. The main goal of process mining is to enhance this practice by offering tools and methods for uncovering various structural aspects, including process, control, data, organizational, and social elements, from event logs. In the industry, this is often referred to as Automated Business Process Discovery (ABPD).

Traditional process mining algorithms assume that the processes being examined remain consistent. However, in reality, processes can undergo changes for a multitude of reasons, such as economic factors, the introduction of new laws (e.g., Sarbanes-Oxley), seasonal variations, or deadline escalations. These changes, collectively known as concept drifts, represent a significant challenge in the field of process mining.

When dealing with the intricate task of analyzing process execution behavior as manifested in event logs, it is crucial to detect and address any concept drift before attempting to construct a process model or applying other process intelligence algorithms. Surprisingly, there has been limited research focused on the detection of concept drift in the context of process mining.

Operational processes often need to adapt to changing circumstances, whether due to new legislation, extreme fluctuations in supply and demand, or seasonal effects. When extracting a process model from event logs, the conventional assumption is that the process at the start of the recorded period remains the same as at the end of the recorded period. However, this assumption often does not hold true due to the phenomenon known as concept drift. While cases are being handled, the underlying process may undergo changes, necessitating an approach to analyze these dynamic shifts.

Our proposed approach has been implemented in ProM3 and validated through an analysis of an evolving process. In the realm of process mining, there are three primary classification techniques: Discovery, Conformity Analysis, and Extension.

Process mining is closely aligned with Business Process Intelligence (BPI), Business Operations Management (BOM), Business Activity Monitoring (BAM), and data workflow mining. The remainder of this paper is structured as follows: related work is discussed in Section II, system design is detailed in Section III, implementation is covered in Section IV, experimental results are presented in Section V, and the paper concludes in Section VI.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

II. RELATED WORK

In the realm of data analysis involving temporally ordered data, concept drift represents a significant concern. Over the past decade, Process mining has emerged as a discipline dedicated to leveraging information system logs for the purpose of mining, analyzing, and improving the process dimension. Notably, there has been limited exploration of the concept drift phenomenon within the domain of process mining.

Recently, an innovative online mechanism for detecting and managing concept drift has been introduced. This mechanism, grounded in abstract interpretation and sequential sampling, combines contemporary data stream learning techniques [1]. Modern process-aware information systems maintain detailed process information as they unfold, offering diverse applications. The term "process mining" encompasses techniques and tools for extracting knowledge, often in the form of models, from this wealth of data. Key players in this field have developed advanced process mining tools, such as Aris PPM and the HP Business Cockpit, which harness available information to yield meaningful insights [2].

Concept drift, in essence, presents a non-stationary learning challenge over time, with discrepancies between training and application data being commonplace in real-world scenarios. In this report, we address the concept drift issue in the context of process mining [3]. We introduce an ensemble method for managing concept drift, which dynamically creates and removes weighted experts in response to performance changes. This method, referred to as dynamic weighted majority (DWM), incorporates four mechanisms to effectively cope with concept drift. These mechanisms involve training online learners, assigning weights based on their performance, removing underperforming learners, and introducing new experts based on the ensemble's global performance [4].

Operational processes often undergo adjustments to adapt to changing conditions, such as shifts in legislation, significant fluctuations in supply and demand, or seasonal influences. While the topic of flexibility is well-explored in the Business Process Management (BPM) domain, contemporary process mining approaches typically assume process stability. When extracting process models from event logs, it is conventionally assumed that the process's characteristics at the beginning of the recorded period persist until the end. However, this assumption often does not hold due to the concept drift phenomenon. During case execution, the process itself may undergo changes, prompting the need for an approach to analyze these second-order dynamics.

Our proposed approach, implemented in ProM3, has been evaluated by analyzing an evolving process [5]. Decision-making in process-aware information systems involves both build-time and run-time decisions. At build-time, idealized process models are designed based on organizational objectives, infrastructure, constraints, and other factors. However, these idealized models can deviate from reality at run-time, particularly when planned activities do not occur within the expected timeframes. Users are then faced with decisions, referred to as escalations, to ensure process completion within the specified timeframe or minimize tardiness. This framework for escalations draws on established principles from the workflow management field [6].

The domain of spam filtering presents a challenging machine learning task, marked by dynamic changes in data distribution and concepts, primarily driven by spammers seeking to bypass spam filters. Lazy learning techniques are well-suited for such dynamically changing contexts. We introduce a case-based system for spam filtering that can adapt dynamically [7].

Processes, while often similar across organizations or within the same organization, may also necessitate local variations. Configurable process models have been advocated to accommodate such variations in a controlled manner [8].

Fluctuations in fuel feeding and fuel inhomogeneity can impact the circulating fluidized bed (CFB) process, leading to efficiency reduction and shortened component lifetimes if control systems fail to compensate for these fluctuations. This paper addresses the challenge of online mass flow prediction, encompassing accurate prediction and explicit detection of abrupt concept drift, along with noise handling mechanisms [9].

Additionally, a classification and lifecycle framework for business process change is introduced to explore the influence of contextual variables on business processes and the need for different responses under varying conditions [10]. Process mining techniques play a crucial role in extracting valuable insights from event logs, including control-flow discovery, which involves the automated construction of process models to depict causal relationships between activities. These insights are pivotal for advancing the next generation of Process-Aware Information Systems (PAISs) as they highlight the disparity between proposed models and reality [11].

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

III. METHODOLOGY

To address the goals set by the framework illustrated in Figure 1, a structured methodology is implemented for the analysis of concept drifts in process mining. This methodology outlines a step-by-step approach to effectively achieve the desired outcomes.

1. **Data Import:** The initial step of the methodology involves the import of actual event logs into a dedicated knowledge base server. This step ensures that the raw data is readily available for analysis and processing. A mathematical representation of the data import process could involve elements such as

   Data Import Function: DIF(data_source, knowledge_base_server)
   Data Import Efficiency: E = (Imported Data Volume) / (Time taken)

2. **Framework Adherence:** Subsequently, the methodology dictates strict adherence to the established framework. This ensures that the entire process is carried out in a systematic and organized manner, in line with the framework's guidelines.This step may involve quantifying the level of adherence to the framework
   Adherence Index: AI = (Number of Framework Steps Followed) / (Total Number of Framework Steps)

3. **Environmental Monitoring:** An essential aspect of the methodology is continuous monitoring of the environment. This step involves observing and recording changes that occur within the process mining environment. These changes are a key component in detecting and understanding concept drifts.To mathematically express environmental changes
   Change Rate Function: CRF(environment_variable) = Δ(environment_variable) / Δ(time)
   Environmental Change Alert Threshold: T_threshold

4. **Event Log Recognition:** The methodology places significant emphasis on the recognition of event logs. This involves categorizing and structuring the event logs in a manner that facilitates subsequent analysis. Suitable visualization techniques are applied to make these event logs comprehensible. In this step, a mathematical representation of event log categorization can be introduced
   Categorization Function: CF(event_log) = Category
   Visualization Effectiveness Score: VES = (Information Captured) / (Complexity of Visualization)

5. **Dynamic Analysis:** An integral part of the methodology is the dynamic analysis of the event logs. This involves utilizing the Promenade VI network to gain insights into the dynamic nature of concept drifts. The dynamic analysis is a crucial step in comprehending how the method is evolving over time.A simple mathematical model for dynamic analysis could involve:
   Event Log Dynamics: EL(t) = f(t), where EL represents event log attributes and f(t) signifies how these attributes change over time.
   Dynamic Analysis Metrics: DAM = (Change Magnitude) / (Change Duration)

6. **Concept Drift Identification:** The methodology is designed to facilitate the identification of concept drifts within the process mining domain. This step is achieved through the collective efforts of the four modules outlined in the framework.To mathematically identify concept drifts, you might consider:
   Concept Drift Score: CDS = (Dissimilarity in Event Logs) / (Time Interval)
   Concept Drift Threshold: T_CDS_threshold

The methodology ensures a structured and systematic approach to concept drift analysis in process mining, allowing for the accurate identification and understanding of changes occurring within the method over time.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

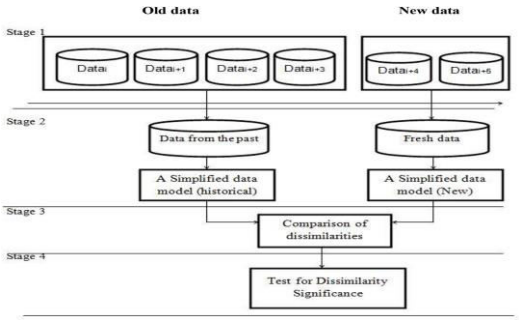

Fig. 1. Frame work for executing concept drifts.

A. Feature Extraction

Event logs contain valuable information about the relationships between activities within a process. These dependencies among activities can be effectively captured and articulated using the follows (or precedes) relationship, often referred to as causal footprints. For any given pair of activities denoted as 'a' and 'b' within the activity set 'A,' and for a trace 't' represented as t(1)t(2)t(3) . . . t(n) belonging to the set A+, we assert that 'b' follows 'a' if and only if, for every 'i' within the range of $1 \leq i \leq n$, where t(i) = a, there exists a 'j' such that $i < j \leq n$, and t(j) = b. In temporal logic notation, this relationship can be represented as _ (a ⇒ (♦b)). Conversely, we declare that 'a' precedes 'b' if, and only if, for all 'j' within the range of $1 \leq j \leq n$, where t(j) = b, there exists an 'i' such that $1 \leq i < j$, and t(i) = a, symbolized as aWb, with 'W' denoting the weak until operator in linear temporal logic notation.

The notions of 'follows' and 'precedes' relationships are not limited to individual traces but can be extended to the entire event log. If 'b' consistently follows 'a' in all the traces contained within an event log, we categorize this relationship as 'b always follows a.' Conversely, if 'b' follows 'a' in only a subset of the traces, we label it as 'b sometimes follows a.' If, in contrast, 'b' never follows 'a' in all traces, we denote this as 'b never follows a.'

Consider an event log labeled as 'L,' comprising three traces defined over the set of activities 'A,' where A = {a, b, c, d, e, f, g, h, i, j, k}. Examining this event log, we can identify the following relationships: 'e' always follows 'a,' 'e' never follows 'b,' and 'b' sometimes follows 'a.' These relationships can be effectively represented within a matrix, where each cell (i, j) is populated with 'A,' 'S,' or 'N,' signifying whether the activity denoted by the column 'j' always, sometimes, or never follows the activity denoted by the row 'i,' respectively.

RE: Relative Entropy (RE) The concept of Relative Entropy (RE) concerning the follows (precedes) relation is described as a function, denoted as f_L_RE: A → R+, defined over the set of activities 'A.' For a specific activity 'x' within 'A,' the RE value with respect to the follows (precedes) relation is computed as the entropy of the RC (Relation Count) metric. In more precise terms, f_L_RE(x) is defined as -pA log2(pA) - pS log2(pS) - pN log2(pN), where pA = cA/|A|, pS = cS/|A|, and pN = cN/|A|, with <cA, cS, cN> being derived from f_L_RC(x). To illustrate, in the case of an event log 'L,' f_L_RE(a) would equate to 0.684, corresponding to f_L_RC(a) = (2, 9, 0), while f_L_RE(i) would be 1.322, corresponding to f_L_RC(i) = (1, 4, 6). When dealing with an event log containing '|A|' activities, this results in a feature vector of dimension '|A|' or 2×'|A|,' contingent on whether one or both of the follows/precedes relations are considered.

WC: Window Count (WC) Within the context of follows (precedes) relations, the Window Count (WC) is defined as a function, f_l,t_WC: A×A → N0, established over the set of activity pairs. Given a trace 't' and a window of size 'l' denoted as $l \in N$, the function f_l,t_WC operates as follows: 'S_l,t(a)' signifies the collection of all subsequences 't(i, i + l - 1)' where 't(i)' = 'a'. More specifically, the function 'F_l,t(a, b)' identifies subsequences within 'S_l,t(a)' where 'b' follows 'a' – this can be denoted as 'F_l,t(a, b) = [$s \in S\_l,t(a) \mid \exists 1 < k \leq |s|$, s(k) = b].' In essence, 'F_l,t(a, b)' represents the set of subsequences that initiate with 'a' and are subsequently followed by 'b' within a window of length 'l.' The WC for the 'b follows a' relationship, f_l,t_WC(a, b), is quantified as the cardinality of 'F_l,t(a, b).'

J Measure The J measure, originally proposed by Smyth and Goodman, is a metric designed to quantify the information content or "goodness" of a rule. Within the context of characterizing the significance of relationships between activities, this metric is employed as a feature. The fundamental concept is rooted in considering the 'b follows a' relationship as a rule: if activity 'a' occurs, it implies that activity 'b' is likely to occur as well. The J measure with respect to follows (precedes) relations is represented as the function f_l,t_J: A×A → R+, operating over the set

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

of activity pairs and within a given window of length 'l' ∈ N. Here, 'p_t(a)' and 'p_t(b)' denote the probabilities of the occurrence of activities 'a' and 'b,' respectively, within a trace 't.' Additionally, 'p_l,t(a, b)' represents the probability that 'b' follows 'a' within a window of length 'l,' computed as 'p_l,t(a, b) = |F_l,t(a, b)|/|S_l,t(a)|.'

These descriptions provide a comprehensive overview of the mathematical concepts involved in Relative Entropy (RE), Window Count (WC), and the J measure with respect to follows (precedes) relations in event log analysis.

IV. EXPERIMENTAL RESULTS

The outcomes of addressing concept drift within the realm of process mining, utilizing event logs, are presented graphically with an X-axis and a Y-axis. This visual representation showcases the dynamics of events as they transition from one process to another, pinpointing the precise moments of change.

In Figure 2, you can observe the visualization of concept drifts. On the X-axis of the chart, we track "Change Detection Over Time," which represents the timeline of changes occurring within the process. This axis chronicles the progression of these changes.

On the Y-axis, we examine "Change Percentage," indicating the extent or magnitude of change during specific time intervals. The Y-axis values demonstrate how the level of change fluctuates, either increasing or decreasing, in response to the different event logs analyzed in the context of process mining.

This visual representation, in the form of a bar chart, provides a comprehensive overview of how the process evolves over time. It aids in the precise identification and understanding of concept drifts within the process, shedding light on how events and patterns change over time.

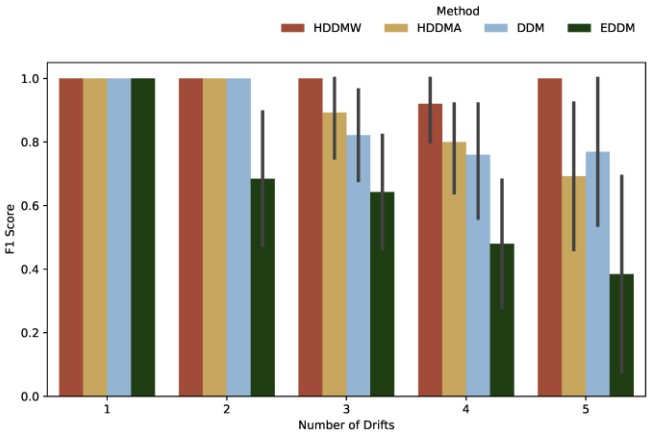

Fig. 2. Visualization of Drifts

V. CONCLUSION

The primary objective of this project is to detect changes within processes extracted from event logs and pinpoint the exact moment of change using a feature extraction mechanism. The project proposes a set of features designed to effectively identify changes in the control flow perspective. Additionally, future work will extend the use of these features to detect changes in both the data and resource perspectives within the process.

In summary, the project's core focus is on enhancing the capability to recognize process changes, and it aims to broaden its application to encompass changes in data and resource aspects in the future. This research and development are valuable in the field of process mining, where the ability to detect and adapt to concept drifts is crucial for maintaining the efficiency and accuracy of process analysis

REFERENCES

[1] Carmona and R. Gavaldà, “Online techniques for dealing with concept drift in process mining,” in Proc. Int. Conf. IDA, 2012, pp. 90–102.

[2] B. F. van Dongen and W. M. P. van der Aalst, “A meta model for process mining data,” in Proc. CAiSE Workshops (EMOI-INTEROP Workshop), vol. 2. 2005, pp. 309–320.

|e-ISSN: 2319-8753, p-ISSN: 2347-6710| www.ijirset.com | Impact Factor: 8.423| A Monthly Peer Reviewed & Referred Journal |

|| Volume 12, Issue 11, November 2023 ||

| DOI:10.15680/IJIRSET.2023.1211015 |

[3] I. Žliobait˙e, "Learning under concept drift: An Overview," CoRR, vol. abs/1010.4784, 2010 [Online]. Available: http://arxiv.org/abs/1010.4784.

[4] J. Z. Kolter and M. A. Maloof, "Dynamic weighted majority: An ensemble method for drifting concepts," J. Mach. Learn. Res., vol. 8, pp. 2755–2790, Jan. 2007.

[5] R.P. Jagadeesh Chandra Bose, Wil M.P. van der Aalst, Indr_Zliobait, and Mykola Pechenizkiy, "Handling Concept Drift in Process Mining", Department of Mathematics and Computer Science, University of Technology, The Netherlands.

[6] W. M. P. van der Aalst, M. Rosemann, and M. Dumas, "Deadline-based escalation in process-aware information systems," Decision Support Syst., vol. 43, no. 2, pp. 492– 511, 2011.

[7] S. J. Delany, P. Cunningham, A. Tsymbal, and L. Coyle, "A case-based technique for tracking concept drift in spam filtering," Knowl. Based Syst., vol. 18, nos. 4–5, pp. 187–195, Aug. 2005.

[8] W. M. P. van der Aalst, N. Lohmann, and M. L. Rosa, "Ensuring correctness during process configuration via partner synthesis," Inf. Syst., vol. 37, no. 6, pp. 574–592, 2012.

[9] M. Pechenizkiy, J. Bakker, I. Žliobait˙e, A. Ivannikov, and T. Kärkkäinen, "Online mass flow prediction in CFB boilers with explicit detection of sudden concept drift," SIGKDD Explorations, vol. 11, no. 2, pp. 109–116, 2009.

[10] K. Ploesser, J. C. Recker, and M. Rosemann, "Towards a classification and lifecycle of business process change," in Proc. BPMDS, vol. 8. 2008, pp. 1–9.

[11] M. Dumas, W. M. P. van der Aalst, and A. H. M. Ter Hofstede, Process Aware Information Systems: Bridging People and Software Through Process Technology. New York, NY, USA: Wiley, 2005

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.423

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN SCIENCE | ENGINEERING | TECHNOLOGY

9940 572 462 6381 907 438 ijirset@gmail.com

www.ijirset.com

Scan to save the contact details