International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

Handling Concept Drift in Data Stream Mining of Event Logs Using Hybrid Optimization Algorithm

Swapna Neerumalla* And L.Ramaparvathy†

Submitted: 10/09/2022 **Accepted:** 20/12/2022

Abstract: Process discovery is a method for attaining process scheme relies on traces exists in the event log. Today, information systems generate the streaming event logs to store their enormous processes. The truncated event log streaming is a challenging problem in process detection since it accuses an incomplete traces, which makes the inaccurate process in a process model. Several conventional techniques have been introduced for retrieving the truncated streaming of event log. This research proposes a method, namely Fractional Improved Invasive Lion Algorithm (FrIILA) for performing the concept drift handling on event log. For that, the event log data is processed under process dimension trimming using bounding model. Moreover, the process mining is carried out using developed FrIILA, which is deliberated by the integration of Fractional calculus (FC) and Improved Invasive Lion Algorithm (IILA). For the incremental data, the same processing is carried out for determining the process discovery. Here, the concept drift detection is carried out using two conditions, such as new event label and max min position of trace. The experimental outcome demonstrates that the devised method achieved better performance based on the replayability and precision of 98.01% and 80.39%.

Keywords: *Fractional calculus, Improved Invasive weed optimization, event log, bounding model, truncated streaming, concept drift detection*

1.Introduction

An event log refers to the collection of various kinds of information in a system. An event log stores various kinds of information, like activity time, activity execution and activity name. These stored information in event log have been mainly utilized in the discovery of problem solving and process assessment. Moreover, event log is utilized to resolve the issues and analyse the implementation processes in a system. Generally, event log contains the process discovery in which lot of information seems to enable the process assessment. Process discovery phase organizes the data of event log in process scheme such that the consumers can effortlessly resolve the processes [1]. Process retrieval refers to the major part in data mining, which has dealt with troubles in various sectors, like medical, fraud, business and advertisement. In order to achieve an accurate process model, the process discovery requires a complete consequence of activities, which is termed as complete traces. Currently, information systems have been generating an event log streaming as their efficiency of processes. However, discovery of process is a difficult process for modelling the streaming as of a truncated event log [2]. The truncated event log streaming generates an imperfect traces as the input information of process detection. Based on the imperfect traces, process detection portrays the incorrect processes in an acquired process model [3].

The growing of information technology offers the creation of huge quantity of high-speed information. The streaming of high speed data exceeds the processing ability of conventional methods [21]. The pre-processing is more significant to copy the instances in a single pass and consider a little set of samples from stream data due to the limited amount of memory space. The main purpose of sampling procedure is to select a serving of data stream, which acts like the whole. The processing of data streaming in imbalanced data is crucial as it seems in several domains, like anomaly detection, weather data forecast, social media excerption and so on. Class imbalance is probable when the instance count with one class is much superior than the others [3]. The class with majority of the data instances are termed as majority classes, whereas the class with minority of the data occurrences are termed as minority classes. For the classification of data streaming, the majority class overpowers the instances as well as disregards the minority classes. Class imbalance is the not only the main issue in the streaming of event log, but the concept drift is the major issue in data streaming [2] [22].

Majority of the classification of data stream representations do not concentrate on the concept drift problem in streaming of imbalanced information, and drift discovery approaches are deliberated based on the statement that the balanced data streams. Concept drift characterizes the modification in distributed occurrences over time, which produces an important problem in the analysis of streaming data [2][24]. Structural decision-making is to discover an best or the maximum acceptable solution for a judgement issue. These decision issues have numerous categories, from everyday operational decisions to

*Research Scholar Department of Computer Science and Engineering
Saveetha School of Engineering
ORCID ID : 0000-0002-4138-8692
† Professor, Department of Computer Science and Engineering,
Saveetha School of Engineering,
Saveetha Institute of Medical and Technical Sciences, chennai
ramaparvathyl.sse@saveetha.com
ORCID:0000-0001-8645-254X
Corresponding Author Email: swapnakiran29@gmail.com

long-term policy business decisions, from an interior single decision to a multi-stage decision or a various organizational decision. Dissimilar decision making responsibilities may have dissimilar features, and soare usually demonstrated in diverse forms or obtained by diverse approaches, which are resolved by diverse decision-making methods. Generally, structural decision issues can be categorized based on their natures [25]. The classic classification is depending on the given structure of problem, that is structured, semi-structured as well as unstructured A structured decision issue can be demonstrated by the classic arithmetical schemes, like linear programming or statistics approaches [23]. The process for attaining the best solution is known as normal solution approaches. Data-driven decision making (D3M) procedures or machine-learning dependent decision-making methods are more appropriate for an unstructured decision issue and for decision making in energetic as well as composite circumstances [3].

The major intention of this research is the development of novel FrIILA for handling the concept drift on event log data. The devised model comprises three phases, such as process dimension trimming, process mining and handling concept drift. Here, the event log data is deliberated as an input, and then the trimming of process dimension is carried out with bounding model. After the dimension trimming is finished, then the process mining is done using devised FrIILA model, which is designed by the assimilation of FC with the IILA. After the completion of process mining, then the concept drift handling is carried out for streaming the process mining. If the concept drift is identified, then the data streaming is done using two conditions based on the threshold value.

The major contribution of the devised model is,

Proposed FrIILA for process mining:In this paper, the process mining is completed using devised FrIILA model, which is designed by the amalgamation of FC and IILA scheme. Moreover, IILA is the integration of IIWO and LA. Moreover, the concept drift handling is done based on the two conditions, such as trace length and class labels.

The organization of this paper is given below. Section 2 explains the review of different traditional process mining and concept drift detection methods, and section 3 presents the projected approach. Section 4 discussed the results and discussion of developed processing model, and finally conclusion is given in section 5.

2. Literature survey

The literature survey of various concept drift detection and process mining techniques are explained in this section. RiyanartoSarno and Kelly RossaSyngkano [1] devised the Coupled Hidden Markov Model (CCMM) for performing the truncated data streaming. Here, the probabilities of states were determined using CHMM. Although, this method was reduced the computational complexity, this method failed to attain the effective classified outcome. In order to improve the prediction performance, Ancy and Paul Raj [2] modelled the ensemble classification and dynamic sampling model for handling the concept drift. Here, the dynamic sampling was carried out using the sample size of optimal reservoir. Moreover, this method attained the better performance using the sample size of optimal reservoir. However, this method was failed to process the real time applications. In order to process with real time applications, Jie Lu *et al.* [3] devised the data driven decision support system for handling the drift in event log data. In this paper, there processes were carried out, namely data streaming, drift detection and understanding and adaptive decision making. The combination of these three processes was made the drift detection process as effective. However, the processing time of this scheme was high. For reducing the processing time, Tobias *et al.* [4] modelled Earth mover’s distance for performing the concept drift. In order to detect the control flow variation among two stochastic data, the post-normalized Levenshtein distance was utilized. Though, the processing time of this scheme was high, the computation overhead of this method was high. To minimize the computation overhead, Hugo de Oliveira *et al.* [5] devised the process mining of event logs using grid-based method. Here, the process mining is done based on the two processes, such as grid process and Petri net unfolding scheme. Moreover, Tabu search model was employed to augment the score of replayability. However, the computation cost of this scheme was high. In order to reduce the computation cost, Martin Prodel*et al.* [6] devised the new hierarchical structuration scheme for handling the huge quantity of information. Moreover, the devised model was designed by integrating Tabu search and Monte Carlo optimization. Although, the effectiveness of devised model was high, it was failed to process with complex data. For attaining the better performance with complex data, HongWei Sun *et al.* [7] modelled the multiple concurrency short loop scheme for performing the process mining. The effectiveness and accuracy of devised model was high. However, this method was failed to process with real world datasets. For performing the process mining in real world datasets, Andrews *et al.* [8] devised the RDB2Log's design for process mining based on semi-automated event log. Although, the computational cost of this scheme was low, this method produced the error value. Hence, the devised model is introduced for resolving these issues.

2.1 Challenges

The challenges of several concept drift detection and process mining techniques are explained in this section.

- The proposed method in [1] is auspicious as this is the first, which employs CHMM to retrieve the streaming of truncated event log. The devised method has difficulty to retrieve event log with new activities [1].
- For retrieving the event log with new activities, an ensemble classification and dynamic sampling model was devised in [2]. There are two challenges are arisen in [2] for the modeling of classification method based on the imbalanced data stream along with concept drift. The major challenge is to utilize the organization scheme while happening the concept drift. The second challenge is to recognize the distribution of imbalanced class and to stop the condition of disregarding the occurrences of smaller class from the data stream.
- For effectively handling the concept drift, data driven decision support system was developed in [3]. An important challenge of utilizing great amount of streaming information acquired from various sources in dissimilar time frames is insecurity. Insecurity in huge-quantity streaming information considers a quantity of dissimilar forms [3].
- For attaining the better performance with huge volume of information, the developed approaches in [4] that permit to methodically estimate a appropriate preprocessing without necessitating field knowledge. Seeing the trace descriptors, like lower-dimensional

descriptors, which integrate supplementary viewpoints are conceivable challenging directions to inspect [4].

- The sampling is a procedure of choosing the occurrences from the repeatedly receiving data streams and evaluating an estimation to indicate the entire stream information. Most of the traditional methods consider the sample occurrences in a stable size. Though, they are unsuitable across the distribution of imbalanced class.

3. Proposed FrIILA for concept drift handling in even log data streaming

This section portrays the devised concept drift handling of data stream mining in event log data using developed FrIILA. The devised model comprises three phases, such as process dimension trimming, process mining and handling concept drift. In the devised model, the event log data is deliberated as an input, and then the trimming of process dimension is done using bounding scheme. Once trim the process dimension, the process mining is carried out using devised FrIILA scheme, which is designed by the assimilation of FC with the IILA. After the completion of process mining, the process discovery is carried out to identify the concept drift. For the incremental data, the same processing, like process dimension trimming, process mining and handling concept drift are carried out for determining the process discovery. Based on the process discovery outcome of both event log data and incremental data, the concept drift is identified. If the concept drift is identified, then the data streaming is carried out using two conditions based on the threshold value.

a) Fitness function

The fitness is calculated to compute the best solution in process mining. Here, the fitness measure with least value is deliberated as optimal solution, and it is assessed using the subsequent equation.

$$Fitness = \frac{G + \left(1 + \tilde{G}\right) + D}{3} \tag{4}$$

where, G signifies replayability score, and D specifies precision, which are indicated as follows,

$$G = \frac{J_o}{M_w} \tag{5}$$

$$\tilde{G} = \frac{F_o}{M_w} \tag{6}$$

$$D = \frac{B_o}{L_w} \tag{7}$$

where, J_o indicates the number of transitions enclosed in solution, M_w states overall transition in solution, F_o depicts the transition counts not completed in solution, B_o describes the transition count completed in table, and L_w denotes complete transition in table.

b) Developed FrIILA for process mining

The processing of developed FrIILA for process mining using event log data is explained in this section. Here, the devised FrIILA is designed by joining the FC, with IILA technique. Moreover, IILA is the integration of IIWO [17] and LA [18]. FC [16] is the mathematical analysis, which is used to resolve the various optimization issues. LA is an optimization model, which considers the hunting aspects of lions. The LA has the ability to solve the complex optimization issues, improved convergence rate and minimized error. Likewise, IIWO algorithm is a population dependent optimization scheme, which is designed by adapting the aspects of weed colony. Moreover, IIWO algorithm is simple and effective. Thus, the devised FrIILA is designed by considering the advantages of FC with the IILA for attaining the better performance. The processing flow of devised FrIILA for process mining is described in the subsequent section.

i) Initialization

Let us initialize P^{male}, P^{female} and P^{nomad} in which P^{male} and its lioness P^{female} contain its pride. The vector components of P^{male}, P^{female} and P^{nomad} are $P^{male}(u)$, and $P^{female}(u)$, $P^{female}(u)$ which indicates the arbitrary integers in highest and least limitations, here u=1,2,.......X where, X states the entire quantity of fractional abundance of optimized terminal members.

ii) Fitness computation

The fitness function of male, female as well as nomad lions are expressed in equation (4), and is specified as F(P^{male}), F(P^{female}) and F(P^{nomad})In order to establish the subsequent steps, let us consider $P^{ref}=F(P^{male})$ and I_h=0, where I_h denotes the generation counter, and is utilized to assess the termination circumstance.

iii) Fertility assessment

In this step, the computation for territorial lioness productiveness and lion productiveness are completed. Furthermore, the fertility evaluation is completed to generate the renewed female lion, and is expressed as $P^{female+}$. Thus, the final updated expression of IILA is given below.

$$p_x^{female+} = \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right)\left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] \tag{8}$$

$$p_x^{female+} - p_x^{female} = \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right)\left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] - p_x^{female} \tag{9}$$

In order to resolve the calculation complexity of IILA method, the FC [16] is applied to the update function of IILA scheme. Then, the expression can be rewritten as,

$$R^{\hbar}\left[p_x^{female+}\right] = \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right)\left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] - p_x^{female} \tag{10}$$

$$\begin{aligned} p_x^{female+} - \hbar p_x^{female} - \frac{1}{2}\hbar p_x^{female-1} - \frac{1}{6}(1-\hbar)p_x^{female-2} - \frac{1}{24}\hbar(1-\hbar)(2-\hbar)p_x^{female-3} = \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right) \\ \left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] - p_x^{female} \end{aligned} \tag{11}$$

$$\begin{aligned} p_x^{female+} = \hbar p_x^{female} + \frac{1}{2}\hbar p_x^{female-1} + \frac{1}{6}(1-\hbar)p_x^{female-2} + \frac{1}{24}\hbar(1-\hbar)(2-\hbar)p_x^{female-3} + \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right) \\ \left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] - p_x^{female} \end{aligned} \tag{12}$$

$$\begin{aligned} p_x^{female+} = (\hbar-1)p_x^{female} + \frac{1}{2}\hbar p_x^{female-1} + \frac{1}{6}(1-\hbar)p_x^{female-2} + \frac{1}{24}\hbar(1-\hbar)(2-\hbar)p_x^{female-3} + \left(\frac{\psi(t)-1}{\psi(t)-2+0.1h_1h_2-0.05h_1}\right) \\ \left[(0.1h_2-0.05)p_x^{male} - \frac{p_{best}}{\psi(t)-1}\left(1-0.1h_1h_2+0.05h_1\right)\right] \end{aligned} \tag{13}$$

where, $\psi(t) = \left(\frac{X-x}{X}\right)^{a} \left(\psi_{initial} - \psi_{final}\right) + \psi_{final} C(t)$

Here, $C(t)$ denotes the chaotic function, h_1 and h_2 varies from [0,1]

iv) Mating

In the mating phase, two processes are carried out, namely mutation as well as crossover. While performing the crossover, four cubs, P^{cubs} are produced, and each cub is produced with respect to the evenly distributed random crossover probability D_z .

$$P^{cubs}(u) = V_u \circ P^{male} + \overline{V}_u \circ P^{female} \tag{14}$$

Where ,V denotes the mask of crossover, and $\overline{V}$ signifies the one's complement of V. P^{cubs} are applied to the mutation in reliable performance with the rate of probability U_y. Moreover, the mutation process generates the new hubs. Once the new hubs are generated, then the gender clustering is carried out, such that the male as well as female cubs, namely P^{mcub} and P^{fcub} are excerpted from cub pool using the primary as well as secondary optimal fitness. After this, their ages are calculated, which is depicted as B_{cub} and it is assigned to zero.

v)Function of cub Growth

The male as well as female cubs, namely P^{mcub} and P^{fcub} are forwarded to the matching mutation with rate V_y. For every update, if the growth of muted cub is better than the previous cub, then the ages of new cubs are incremented to one till the muted cub interchanges old cub.

vi)Territorial defense

It is considered as a primary lion operator for investigating the search space in wider mode. Moreover, this operator is sequenced with respect to the update of creating pride, survival fight as well as Nomad coalition. Thus, the winner nomad is selected based on the subsequent expression, which is given by,

$$Y\left(P^{e\_nomad}\right) < Y\left(P^{male}\right) \tag{15}$$

$$Y\left(P^{e\_nomad}\right) < Y\left(P^{m\_cub}\right) \tag{16}$$

$$Y\left(P^{e\_nomad}\right) < Y\left(P^{f\_cub}\right) \tag{17}$$

In case, P^{male} is beaten in territorial defence, then the pride is renewed by substituting P^{male} with conquer nomad coalition $P^{e-nomad}$, and it is renewed by selecting only one P^{nomad}. This process selects P_1^{nomad}, when $W_1^{nomad} \geq e$ is fulfilled, else select P_2^{nomad} and e signifies the exponential unity. Moreover, W_1^{nomad} is assessed by,

$$W_1^{nomad} = \exp\left(\frac{n_1}{\max(n_1, n_2)}\right) \max \frac{\left(Y\left(P_1^{nomad}\right), Y\left(P_2^{nomad}\right)\right)}{Y\left(P_1^{nomad}\right)} \tag{18}$$

where, Euclidean distance amongst the pair $\left(P_1^{nomad}, P^{male}\right)$ is stated as n_1, and n_2 signifies Euclidean distance among the pair $\left(P_2^{nomad}, P^{male}\right)$. If defence outcome is zero, then P^{male} and $Y(P^{male})$ are stored from fertility computation, and the process is continual.

vii) Territorial takeover

If the condition $N_{cub} \geq N_{\max}$ is satisfies, then the territorial takeover is carried out, else the cub growth function is repetitive. This process offers territory to P^{mcub} and P^{fcub} after they grown-up and progresses as tougher than P^{male} and P^{female} .

viii) Evaluation of feasibility

The better solution is calculated with fitness value by equation (4), and if a novel solution is superior than the preceding solution, then renew an preceding value with new value.

ix) Termination condition

The above mentioned procedures are continued until better solution is acquired. The pseudo code of devised FrIILA for process mining is characterized in algorithm 1.

Algorithm 1. Pseudo code of developed FrIILA

```
Pseudo code of devised IILA
Input: Pride
Output: Optimal solution
Begin
Pride initialization
Approximate the fitness value with equation (4)
Assess the fertility by equation (8)
    Mating is completed using equation (14).
Process cub growth function
Winner nomad is nominated with equation (15), (16) and (17)
Assess territorial takeover
Validate the solution feasibility
Renew best solution
Stop
```

Based on the developed FrIILA scheme, the process mining is carried out. After the completion of process mining, the concept drift handling is performed for both event log data and incremental data.

4. Concept drift handling in streaming event log data

One of the major issues in process mining is concept drift of trimming data. Thus, the effective handling of concept drift is a challenging task [9][10][11]. Process mining helps to retrieve, observe and enhance the real processes by excerpting the intelligence from event logs exist in recent information systems. The beginning point of process mining is an event log data. Every event exist in the log represents the activity, which is based on the specific case. The events comes under the specific case are arranged and realized as one run of the process. Generally, some of the excess event attributes are stored in the event logs. Moreover, the process mining approaches utilized the attributes, such as resource, activity execution, and event timestamp and data components with the event. The major cause of affecting data stream is concept drift, which causes due to the varying relation among system response and recorded tuple. The disregarding of concept drift may weaken the effectiveness of algorithm and it has the ability to demonstrate the most modern concepts in data. For performing the validation of any learning process, the concept drift is significant since it modifies the ratio among observed data as well as response. The implementation of a concept-drift adaptation scheme is an essential task as various kinds of concept drift are conceivable and various adaptations are to be processed in response to them.

For the processing of machine learning and data mining approaches, the second order dynamics are considered as concept drift, which can be resolved in both supervised as well as unsupervised settings. Concept drift is a significant task in various computer applications. There are several traditional concept drift handling strategies have been invented by the researchers. However, the conventional concept drift handling approaches have simple structure with varying variables rather than transforms to difficult artifacts, like choices, cancelation, loops, concurrency and so on. In order to handle the concept drift, the subsequent major issues have been recognized.

- Change (Point) Recognition: The most significant and basic issue is to recognize the concept drift or process variation takes place in the processes. In case, the concept drift is identified, then the next stage is to recognize the time duration at which variations have taken place.
- Change Characterization and Localization: Once the varying point is recognized, then the next step is to identify the nature of variation and area of variation. Moreover, the recognition of the nature of variation is not a simple task, which involves the recognition of exact variation and the perspective variation, like data, control-flow, resource and so on.
- Unravel Process Development: Having recognized, characterized and localized the variations, then it is essential to place all of these in viewpoint. There is a necessity for methods/tools, which exploit as well as relate these discoveries. Unscrambling the progression of a process should produce in the detection of variation process.

In this paper, the concept drift handling is carried out using two conditions when the concept drift is detected. The concept drift handling is performed for both event log data as well as incremental data. Here, the concept drift is identified by performing the process dimension trimming, process mining and process discovery on both event log data and incremental data. On comparing the process discovery outcome of both event log data and incremental data, the concept drift is identified. Moreover, the data streaming of concept drift handling is done using 2 conditions. The concept drift handling is carried out based on 2 conditions are given by,

a) New class labels
b) Trace length, i.e., max and min values.

a) For incremental data, if the labels B is greater than label limit N, then the data is considered as a drift data, and then redo the step-1 to step-3 from section 3.1 is repeated for the incremental data, otherwise no change is done in the data.

If $B > N$; where B signifies the labels and N denotes the label limit

Redo the process mining using steps 1 to 3 from section 3.1 for the incremental data

Else

No change

b) For incremental data, if min-max value in trace length L is greater than threshold T, then the data is considered as a drift data, and then redo the process mining for the new drift data using the step-1 to step-3 from section 3.1. Here, the threshold value is set as (10/200), which is set as default value. Moreover, the minmax value is varied depending on the dataset.

If $L > T$; where L signifies the min-max values in trace length and T denotes the threshold value

Redo the process mining using steps 1 to 3 from section 3.1 for the incremental data

Else

No change

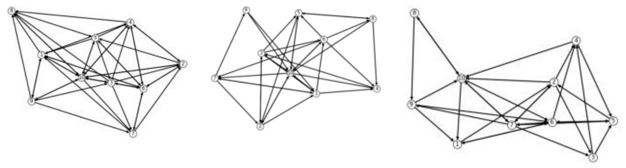
If the concept drift is identified, then redo the process mining is for the new data, otherwise no need of redoing the process mining.

5. Results and discussion

This section illustrates the results and discussion of devised FrIILA for process mining in order to perform the concept drift handling of event logs. Moreover, the setup of experiment, dataset details, performance parameters and comparative techniques are also discussed in this section.

5.1 Node Sample of experimentation

This section explains the experimental sample of devised scheme based on three datasets. Fig. 1. a) illustrates the experimental sample of devised scheme with BPIC12 (Business Process Intelligence Challenge 12), Fig. 1. b) demonstrates the process sample of devised model with BPIC17_f & Fig. 1.c) indicates the process sample of devised model with BPIC15_4f. From the sample image, the numerals 1, 2 and so on indicate the number of nodes and the line arrow deliberates the arc representation.



(a) (b) (c)

Figure 1. Process samples devised using BPIC12, BPIC17_f, BPIC15_4f

5.2 Performance metrics

The metrics employed for the investigation of devised model is replayability score as well as precision.

Replayability score: It is a parameter, which is calculated by the ratio of count of transitions completed in solution to the overall transition in solution, and is expressed as in equation (5).

Precision: It is expressed as the proportion of count of transitions completed in table to the overall transition in table, and is expressed in equation (7).

5.3 Comparative techniques

The existing techniques utilized for the experimentation of devised FrIILA model for process mining is Coupled Hidden Markov Model [1], Ensemble classification model [2], Data-driven decision support model [3] and IILA+Coupled Hidden Markov Model (CHMM).

5.4 Comparative analysis

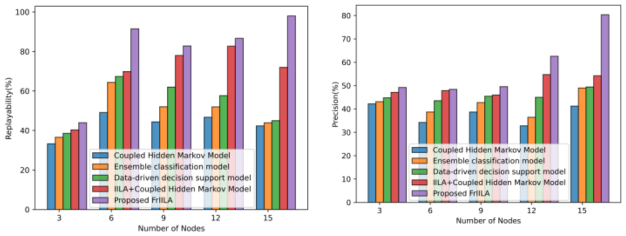
This section depicts the comparative assessment of devised FrIILA based on using three datasets with number of nodes as well as arcs.

5.4.1 Comparative assessment using BPIC12

This section explains the comparative assessment of developed FrIILA scheme by adjusting the number of nodes and number of arcs based on the evaluation metrics using BPIC12.

(i) Assessment by adjusting the number of nodes

Figure 7 a) shows the comparative assessment of devised model by adjusting the number of nodes based on replayability. Here, the replayability attained by the developed scheme is 98.01%, whereas the conventional approaches, such as Coupled Hidden Markov Model attained the replayability of 42.31%, Ensemble classification model attained the replayability of 43.91%, Data-driven decision support model attained the replayability of 45.00%, and IILA+Coupled Hidden Markov Model attained the replayability of 71.97% for the node count is 15. Figure 7 b) demonstrates the assessment of precision based on node count. Here, the precision of 62.56% is attained by the devised scheme when the node count is 12, whereas the precision of 32.77%, 36.43%, 44.99% and 54.74% is attained by the devised scheme when the node count is 12.

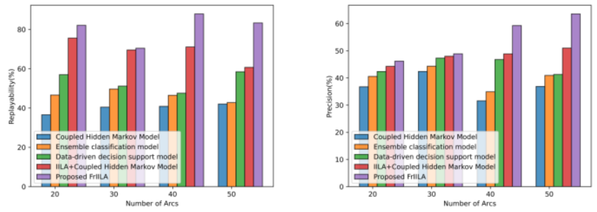


(a) (b)

Figure 2. Comparative assessment of devised scheme using BPIC12 based on number of nodes a) Replayability b) Precision

(ii) Assessment by adjusting the number of arcs

Fig. 3.(a) illustrates the assessment of devised scheme by changing the number of arcs based on the replayability. Here, the replayability value achieved by the devised scheme is 83.29%, whereas the traditional techniques attained the replayability value of 42.03%, 42.76%, 58.45% and 60.69% for the number of arc is 50. Fig. 3.(b) demonstrates the assessment of devised scheme with respect to precision. When the number of arc is 50, then the precision value achieved by the devised scheme is 59.29%, whereas the conventional techniques, such as Coupled Hidden Markov Model, Ensemble classification model, Data-driven decision support model and IILA+CHMM attained the precision values of 31.58%, 34.91%, 46.76% and 48.85%.



(a) **(b)**

Figure 3. Comparative assessment of devised scheme based number of arcs a) Replayability, b) Precision

5.5 Comparative discussion

Table I explains the comparative discussion of devised FrIILA for concept drift handling in even log data streaming. From the Table I, the developed model attained the better performance based on replayability and precision using BPIC12 of 98.01% and 80.39%, whereas the conventional techniques achieved the replayability of 42.31%, 43.91%, 45% and 71.97% , and then the precision values of 41.23%, 49%, 49.42% and 54.22%, correspondingly. In this research, the assessment of devised scheme is done using three setups such as BPIC12, BPIC17-f and BPIC15-4f. From these, BPIC12 attained the better values of replayability and precision due to the effectives of devised FrIILA scheme for process mining. The developed FrIILA is designed by the combination of FC with the IIWO and LA. The optimization algorithm, such as IIWO and LA has better searching ability and the computational complexity of optimization can be diminished by involving the FC such that the combined function of FC, IIWO and LAS provides the better process mining outcome.

Setup	Variations	Metrics	(a)	(b)	(c)	(d)	(e)
BPIC12	Number of nodes	Replayability (%)	42.31	43.91	45.00	71.97	98.01
		Precision (%)	41.23	49.00	49.42	54.22	80.39
	Number of arcs	Replayability (%)	42.03	42.76	58.45	60.69	83.29
		Precision (%)	36.85	40.93	41.29	51.01	63.58
BPIC17-f	Number of nodes	Replayability (%)	39.79	40.33	41.50	61.24	86.38
		Precision (%)	40.38	41.00	59.51	60.98	64.10
	Number of arcs	Replayability (%)	37.09	43.55	46.84	78.92	94.60
		Precision (%)	45.15	49.26	60.54	63.66	66.36
BPIC15-4f	Number of nodes	Replayability (%)	41.01	46.56	47.03	82.09	84.01
		Precision (%)	46.80	48.10	49.15	87.89	91.35
	Number of arcs	Replayability (%)	36.27	37.53	46.68	69.67	75.69
		Precision (%)	32.82	35.52	42.77	80.53	82.18

Table I. Comparative discussion

(a) Coupled Hidden Markov Model
(b) Ensemble classification model
(c) Data-driven decision support model
(d) IILA+Coupled Hidden Markov Model
(e) Proposed FrIILA

6. Conclusion

This research proposes a method, namely FrIILA for performing the concept drift handling on event log. For that, the event log data is processed under process dimension trimming using bounding model. In this paper, the process mining is done using devised FrIILA model, which is designed by the amalgamation of FC and IILA scheme. The concept drift handling is carried out using two conditions when the concept drift is detected. The concept drift handling is performed for both event log data as well as incremental data. Here, the concept drift is identified by performing the process dimension trimming, process mining and process discovery on both event log data and incremental data. On comparing the process discovery outcome of both event log data and incremental data, the concept drift is identified. Moreover, the data streaming of concept drift handling is done using 2 conditions based on trace length and class labels. Based on these two conditions, the concept drift is identified. The experimental result demonstrates that the devised method attained better performance based on the replayability of and precision of 98.01% and 80.39%. However, the developed process mining model can be further protracted by adapting any other effective optimized methods.

Acknowledgements

This research was supported/partially supported by [Vignana Bharathi Institute Of Technology, Gatkhesar, Telangana]. We thank our colleagues from [Vignana Bharathi Institute Of Technology] who provided insight and expertise that greatly assisted the research, although they may not agree with all of the interpretations/conclusions of this paper. We thank [Dr.Pradeep Kumar Mallick, Associate Professor, School Of Computer Engineering, Kalinga Institute Of Industrial Technology, Deemed to be University, Bhubaneswar] for assistance with [analysis

techniques, methodology] that greatly improved the manuscript.

Author contributions

Name1: Swapna Surname1: Neerumalla Conceptualization, Methodology, Software, **Name2: Ramaparvathy Surname2: L** Data curation, Writing-Original draft preparation, Software, Validation.

Conflicts of interest

The authors declare no conflicts of interest.

References

[1] Bose, R.J.C., Van Der Aalst, W.M., Zˇliobaite˙, I. and Pechenizkiy, M., "Dealing with concept drifts in process mining", IEEE transactions on neural networks and learning systems, vol.25, no.1, pp.154-171, 2013.

[2] Tulilaulu, A., Paalasmaa, J., Waris, M. and Toivonen, H., "Sleep musicalization: Automatic music composition from sleep measurements", In International Symposium on Intelligent Data Analysis, pp. 392-403, 2012.

[3] Sarno, R. and Sungkono, K.R., "Recovering Truncated Streaming Event Log Using Coupled Hidden Markov Model", International Journal of Pattern Recognition and Artificial Intelligence, vol.34, no.04, pp.2059012, 2020.

[4] Martjushev, J., Bose, R.P. and Van Der Aalst, W.M., "Change point detection and dealing with gradual and multi-order dynamics in process mining", In International Conference on Business Informatics Research, pp. 161-178, 2015.

[5] Lu, J., Liu, A., Song, Y. and Zhang, G., "Data-driven decision support under concept drift in streamed big data", Complex & Intelligent Systems, vol.6, no.1, pp.157-163, 2020.

[6] Ancy, S. and Paulraj, D., "Handling imbalanced data with concept drift by ap- plying dynamic sampling and ensemble classification model", Computer Communica- tions, vol.153, pp.553-560, 2020.

[7] De Oliveira, H., Augusto, V., Jouaneton, B., Lamarsalle, L., Prodel, M. and Xie, X., "An optimization-based process mining approach for explainable classification of timed event logs", In proceedings of 2020 IEEE 16th International Conference on Automation Science and Engineering (CASE), pp.43-48, August 2020.

[8] W. M. P. van der Aalst, "Process Mining: Discovery, Conformance and Enhance- ment of Business Processes", 2011.

[9] Van Dongen BF, de Medeiros AK, Verbeek HM, Weijters AJ, van Der Aalst WM., "The ProM framework: A new era in process mining tool support", In proceedings of International conference on application and theory of petri nets, pp.444-454, June 2005.

[10] Gu¨nther CW, Van Der Aalst WM., "Fuzzy mining–adaptive process simplification based on multi-perspective metrics", InInternational conference on business process management, pp. 328-343, September 2007.

[11] Brockhoff, T., Uysal, M.S. and van der Aalst, W.M., "Time-aware concept drift detection using the earth mover's distance", In 2020 2nd International Conference on Process Mining (ICPM), pp. 33-40, 2020.

[12] De Oliveira H, Augusto V, Jouaneton B, Lamarsalle L, Prodel M, Xie X., "Op- timal process mining of timed event logs", Information Sciences, vol.528, pp.58-78, August 2020.

[13] Prodel M, Augusto V, Jouaneton B, Lamarsalle L, Xie X., "Optimal process mining for large and complex event logs", IEEE Transactions on Automation Science and Engineering, vol.15, no.33, pp.1309-25, January 2018.

[14] Sun H, Liu W, Qi L, Du Y, Ren X, Liu X., "A process mining algorithm to mixed multiple-concurrency short-loop structures", Information Sciences, vol.542, pp.453-75, January 2021.

[15] Andrews R, van Dun CG, Wynn MT, Kratsch W, R¨oglinger MK, terHofstede AH., "Quality-informed semi-automated event log generation for process mining", Decision Support Systems, vol.132, pp.113265, May 2020.

[16] Misaghi M, Yaghoobi M., "Improved invasive weed optimization algorithm (IWO) based on chaos theory for optimal design of PID controller", Journal of Computational Design and Engineering, vol.6, no.3, pp.284-95, July 2019.

[17] Rajakumar BR., "Lion algorithm for standard and large scale bilinear system identification: a global optimization based on Lion's social behavior", In proceedings of 2014 IEEE congress on evolutionary computation (CEC), pp.2116-2123, July 2014.

[18] Bhaladhare, P.R. and Jinwala, D.C, "A clustering approach for the-diversity model in privacy preserving data mining using fractional calculus-bacterial foraging optimization algorithm," Advances in Computer Engineering, 2014.

[19] Zheng, C., Wen, L. and Wang, J., "Detecting process concept drifts from event logs. In OTM Confederated International Conferences" On the Move to Meaningful Internet Systems", pp. 524-542, 2017.