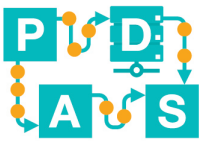


RWTH AACHEN UNIVERSITY

Model Repair by Incorporating Negative Instances In Process Enhancement

Master Thesis

Author : **Kefang Ding**

Supervisor : Dr. Sebastiaan J. van Zelst

Examiners : Prof. Wil M.P. van der Aalst
Prof. Thomas Rose

Registration date : 2018-11-15

Submission date : 2019-04-08

This work is submitted to the institute

PADS RWTH University

Acknowledgments

The acknowledgments and the people to thank go here, don't forget to include your project advice.

Abstract

Big data projects have becomes a normal part of doing business, which raises the interest and application of process mining in organizations. Process mining combines data analysis with modeling, controlling and improving business processes, such that it bridges the gap of data mining on big data and business process management.

Process enhancement, as one of the main focuses in process mining, improves the existing processes according to actual execution event logs . It enables continuous improvement on business performance in organizations. However, most of the enhancement techniques only consider the positive instances which are execution sequences but lead to high business performance outcome. Therefore, the improved models tend to have a bias without the use of negative instances.

This thesis provides a novel strategy to incorporate negative information on process enhancement. Firstly, the directly-follows relations of business activities are extracted from the given existing reference process model, positive and negative instances of actual event log. Next, those relations are balanced and transformed into process model of Petri net by Inductive Miner. At end, long-term dependency on Petri net is further analyzed and added to block negative instances on the execution, in order to provide a preciser model.

Experiments for our implementation are conducted into scientific platform of KNIME. The results show the ability of our methods to provide better model with comparison to selected process enhancement techniques.

Chapter 1

Introduction

Process mining is a relatively new discipline that has emerged from the need to bridge the gap of data mining and business process management. The objective of process mining is to support the analysis of business process, provide valuable insights on processes and further improve the business execution. According to [15], techniques of process mining are divided into three categories: process discovery, conformance checking and process enhancement. Process discovery techniques focus on deriving process models from event logs of the information system, allowing the vision into the real business process. Conformance checking analyzes the deviations between an referenced process model and observed behaviors driven from its execution. Enhancement adapts and improves existing process models by extending the model with additional data perspectives or repairing the existing model to accurately reflect observed behaviors.

Due to the increasing availability of detailed event logs of information systems, process mining techniques have recently enabled wider applications of process mining in organizations around the world[15]. After applying process discovery in organizations, a process model is fixed in information system to guide the execution of business. However,in real life, business processes often encounter exceptional situations where it is necessary to execute process differing from the predefined model. To reflect the reality, the organizations need to adapt the existing process model. Basically, one can apply process discovery techniques again to obtain a new model from event log. However, due to the facts, (1) the cost of rediscovery, and (2) the discovered model tend to have less similarity with the original model[8]. As shown in [8], there is a need to change an existing model similar to the original model while replaying the current process execution. Here comes the model repair.

Model repair belongs to process enhancement and stands between process discovery and conformance checking. It analyzes the workflow deviations between event log and process model, and fix the deviations mainly by adding sub processes on the model. As known, business in organizations is goal-oriented and aims to have high performance according to a set of Key Performance Indicators(KPIs), for example, average conversion time for the sales, payment error rate for the finance. However, there are few researches on applying the process mining with consideration of performance[10]. [10] points out the rare contributions like [6] to combine performance into process mining. Deviations are firstly analyzed to determine if they have a positive impact on the process performance. Model repair techniques in [9] are applied into traces with positive deviations.

However, the current repair methods have some limits. Model repair fixes the model by

adding subprocesses, silent transitions or loops, it guarantees the model fitness but overgeneralizes the model, such that it allows more behaviors than expected. On the other hand, it increases the model complexity. Even the performance is considered in [6], but only deviations in positive is used to add subprocesses, the negative information is ignored, which disables the possibility to block negative behaviors from model. A motivation example is listed to describe those limits.

1.1 Motivation Example

This section describes some situations where current repair techniques can't handle properly. For the sake of understanding, some examples are extracted from the registration procedure of thesis project at one German university to illustrate those situations.

The main activities for the registration process include topic selection, make proposal, meeting with supervisor to discuss the topic, and finish course requirements. After finishing all of those activities, the formal registration is enabled and the procedure comes to end. Figure 1.1 shows the original process in Petri net. The activities are modeled by the corresponding ***transitions*** which is represented by a square. Transitions are connected through a circle called ***place***. Transitions and places build the static structure of Petri net. ***Tokens*** in the black dot are put in the initial places and represent the dynamic state of the model.

The model in Figure 1.1 is in an initial state where only one token is at the start place to enable the first transition begin thesis. After firing *begin thesis*, the token at the initial place is consumed while two new token are generated in the output places. In this way, activity *finish course requirements* can be executed in concurrency with the other activities except for the *register thesis*. When multiple activities have the same input place, all of them are enabled but only one of them can be fired and executed, namely, they are exclusive to each other. As shown in the figure, *select existing topics* and *create new topics* are exclusive, and only one of them can be triggered. When a transition have multiple input places, it can be triggered with condition that all input places hold at least a token. *Register thesis* is enabled only after *finish course requirements* and getting the *approve* done.

In the following part, those situations are introduced to demonstrate the shortcomings of current techniques.

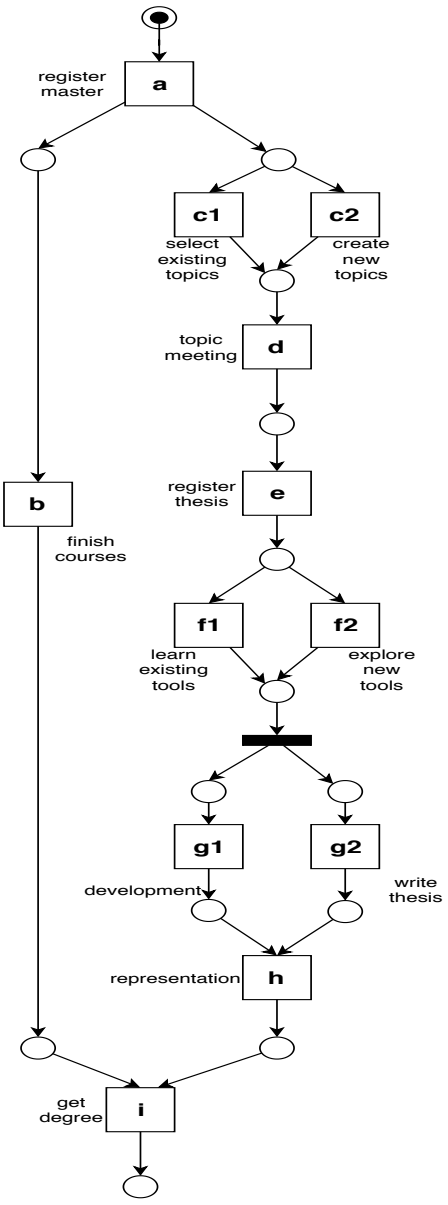


Figure 1.1: original registration process

1.1.1 Situation 1: repair model with unfit traces

Following the given model, in order to address the work between *make proposal* and before the activity *topic meeting*,

two exclusive activities *prepara carefully* and *prepara casually* are executed in the real life and constitutes the actual event log L_1 listed below. With those two activities, it makes the whole procedure more efficient and clear and the traces with them are considered positive. For convenience, alphabet characters are used to represent the corresponding activities and annotated in the model. **e1, e2** represent the activities *prepare carefully* and *prepare casually.*
Event Log L_1 –

$$\begin{aligned} Positive : \{ &< a, b, c1, d, \mathbf{e1}, f, g1, g2, h, i >^{50}, \\ &< a, b, c2, d, \mathbf{e2}, f, g2, g1, h, i >^{50} \} \end{aligned}$$

Because the repair techniques in [9] don't distinguish the performance of event log, to make the model enforcing the positive performance, only positive instances are employed to repair the model. Firstly, the deviations of the existing model in Figure 1.1 and the event log L_1 is computed. After computation of deviations, each deviation has the same start and end place and two deviations appear at the same position of model. When repairing this model, each subprocess has one place as its start and end place, which forms a loop in the model. If there is only one such subprocess, the subprocess will be added in an sequence in the model. This leads to a higher precision. Yet the algorithm does not discover orderings between different subprocesses at overlapping locations. So subprocesses are kept in loop form.

The repaired model is shown in Figure 1.2a, where the two additional activities are added in the form of loop. Compared to the model in Figure 1.2b where the two extra activities are shown in sequence with others, the repaired model in Figure 1.2a has less precision.

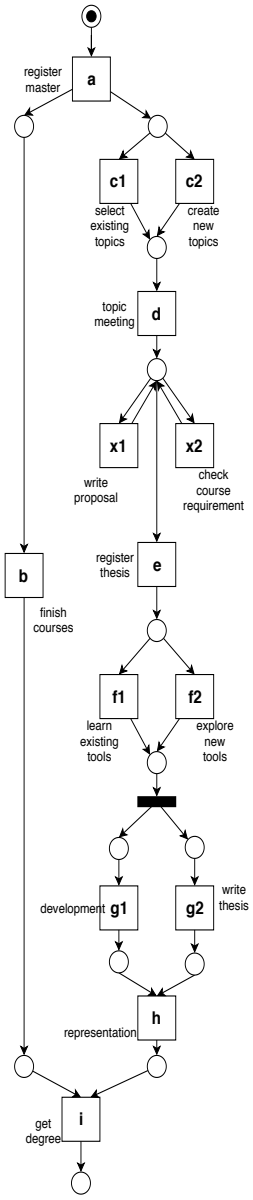
The repair algorithm in [6] builds upon [9] and considers the performance of event log. However, the repaired model is the same as the one in Figure 1.2a. The reasons are: (1) there is no deviation from negative factors. (2) positive deviations are used in the same way like [9]. As a conclusion, the repair techniques in [6] can't deal with situation 1 properly, either.

1.1.2 Situation 2: repair model with fit traces

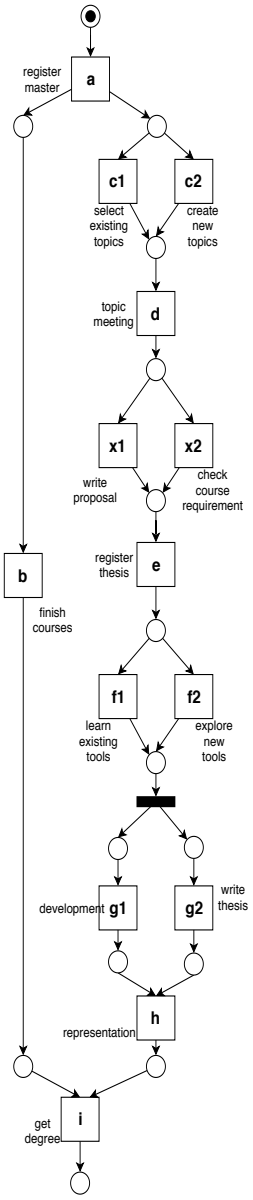
This situation describes the existing problem in the current methods that fit traces with negative performance outcomes can not be used to repair model. Given an actual event log L_2, when activity *finish course requirements* is fired after *begin thesis* and before the topic selection part, it reduces the pressure of master thesis and has a positive outcome. Else, the negative outcomes are given. *Event Log* L_2 –

$$\begin{aligned} Positive : \{ &< a, \mathbf{b}, c1, d, f, g2, g1, h, i >^{50}, \\ &< a, \mathbf{b}, c2, d, f, g1, g2, h >^{50} \} \\ Negative : \{ &< a, c1, d, f, g2, g1, \mathbf{b}, h, i >^{50}, \\ &< a, c1, \mathbf{b}, d, f, g1, g2, h, i >^{50}, \} \end{aligned}$$

Compared to the model, the event log L_2 contains no deviation. When we apply the techniques in [9] and [6] to repair the model, the model keeps untouched due to no deviation. Apparently, those two methods can't incorporate the negative information in fit traces. When we expect a model which enforce the positive instances and avoid the negative instance as the model in Figure 1.3a, the current methods fail our need.

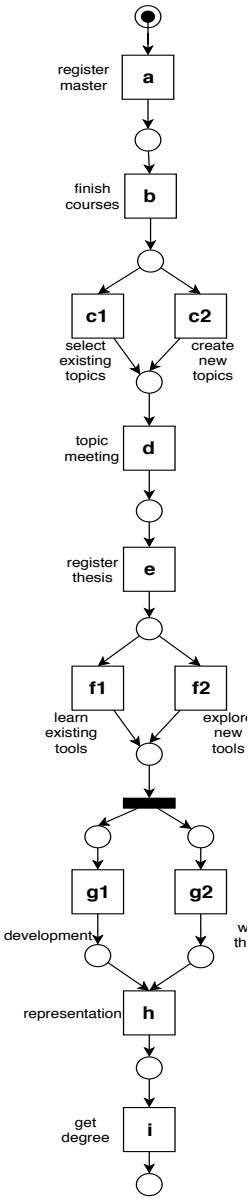


(a) repaired model with additional activities

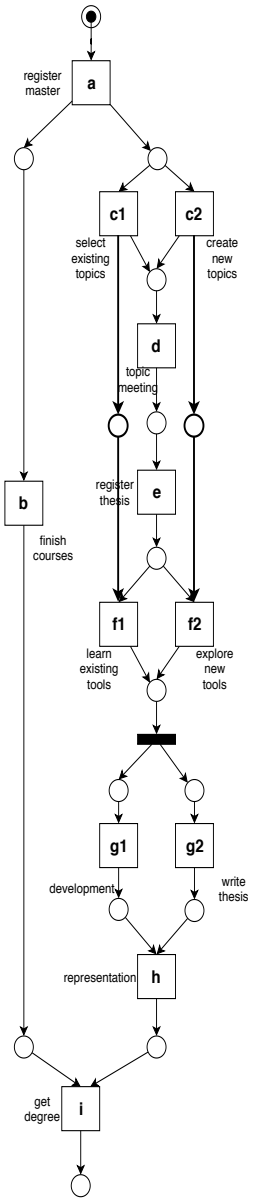


(b) expected model with additional activities

Figure 1.2: example for situation 1



(a) expected model with order change



(b) model with long-term dependency

Figure 1.3: example for situation 2 and 3

1.1.3 Situation 3: detect long-term dependency

This part introduces a problem which causes a lower precision in process model. It is the inability in current methods to detect the long-term dependency in the Petri net. The long-term dependency describes the phenomenon that one execution choice decides the execution of activities that do not follow directly. Due to the long distance of this dependency, current methods can not detect it and improve the precision by adding long-term dependency on model. An event log L_3 is given in the following. By using time consumption as one KPI, if the total sum goes over one threshold, we mark this trace as negative, else as positive. Since *create new topics* usually demands new knowledge rather than checking the existing tools. So if student chooses to learn existing model, it's possibly not useful and wastes time. In the other case, if we select existing topics with existing background, it saves time when we directly learn the existing tools. According to this performance standard, we classified those event traces. *Event Log* L_3 –

$$\begin{aligned} Positive : \{ &< a, b, \mathbf{c1}, d, e, \mathbf{f1}, g1, g2, h, i >^{50}, \\ &< a, b, \mathbf{c2}, d, e, \mathbf{f2}, g2, g1, h, i >^{50} \} \\ Negative : \{ &< a, b, \mathbf{c1}, d, e, \mathbf{f2}, g2, g1, h, i >^{50}, \\ &< a, b, \mathbf{c2}, d, e, \mathbf{f1}, g1, g2, h, i >^{50} \} \end{aligned}$$

There is no deviations of the model and event log L_3 according to the algorithms in [9] and [6]. Therefore, the original model stays the same and allows the execution of negative instances. After checking the model and log, two long-term dependency have significant evidence. Transition ***c1*** decides ***f1*** while ***c2*** decides ***f2***. After addressing long-term dependency like the model in Figure 1.3b by connecting transitions to extra places, negative instances are blocked and the model has a higher precision.

Clearly, the use of negative information can bring significant benefits, e.g, enable a controlled generalization of a process model: the patterns to generalize should never include negative instances. The demand to improve current repair model techniques with incorporating negative instances appears. In the next section, the demand is analyzed and defined in a formal way.

1.2 Research Problem And Questions

We analyze the current model repair methods, and give the formal definitions.

Definition 1.1. Given an input of one existing process model M, an event log L with performance outcomes, how to improve current process enhancement techniques by incorporating negative information, and generate a process model to enforce the positive instances while blocking the negative instance, with condition that the generated model should be as similar to the original model as possible Therefore, the repaired model provides a better way to understand and execute the real business process compared to the original model.

An event log and an existing model are given as the input. According to some predefined KPIs, each trace in event log is classified into positive or negative. After applying repair techniques in the black box, the model should be improved to enforce the positive instances while disallowing negative instance.

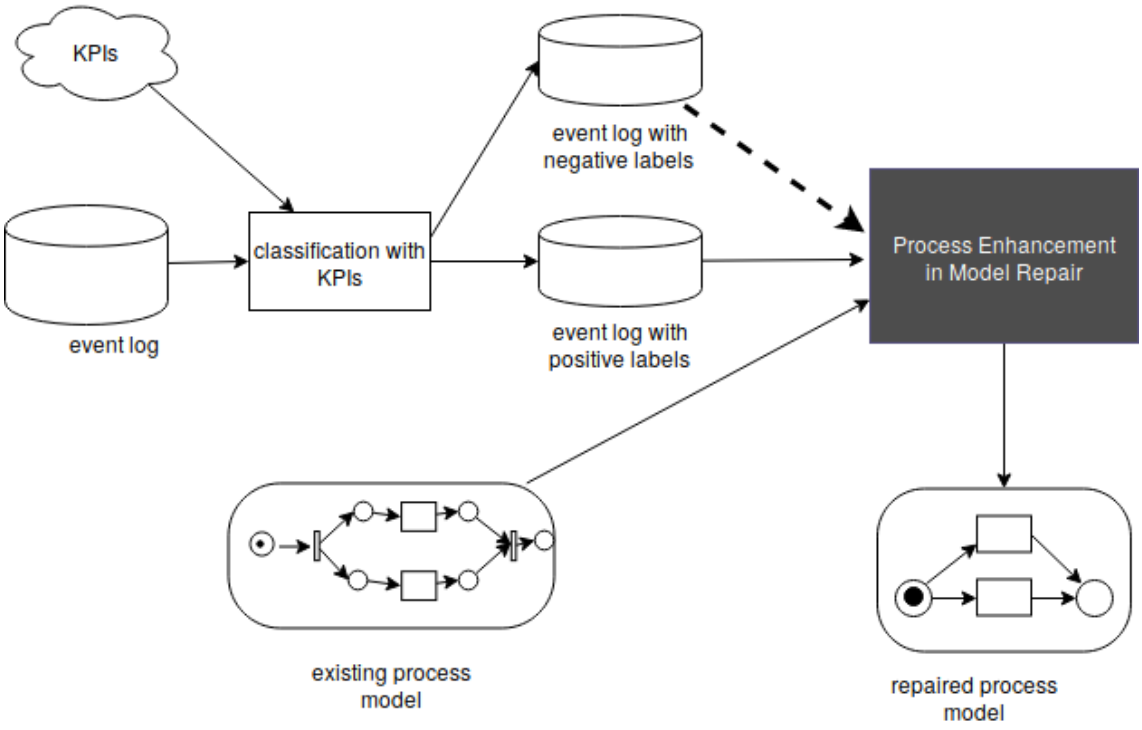


Figure 1.4: The problem description

This paper tries to provide a solution for the black box. Our idea is to analyze the positive and negative impact on process performance of each trace. It balances the existing model, positive traces and negative traces on directly-follows relation, in order to incorporate all the factors on model generation. Later, the directly-follows relation is used to create process model by Inductive Miner. What's more, the impact of the existing model, positive and negative instances are parameterized by weights, to allow more flexibility of the generated model.

1.3 Outline

The reminder is organized in the following order. Section 2 recalls the basic notions on process mining and list the preliminary to solve the problem. The next section lists our methods are introduced and formal definitions are given. In Implementation Section, the details of algorithms are given. Later, we evaluate our methods with simulated data and real data respectively and list the results. Subsequently, the discussion on this paper is presented. At last section, a conclusion is drawn on the paper.

Chapter 2

Related Work

To update an existing process model in organizations, there are two strategies, rediscovery and process enhancement. Process rediscovery applies the discovery techniques on the actual event log to mine a new model. Process enhancement improves the model based on not only the actual event log but also the existing model.

Process discovery has been intensively researched in the past two decades and many algorithms have been proposed[16]. Directly-follows[18, 12] methods investigates the activities order in the traces and extracts higher relations which are used to build process models. State-based methods like [1, 4] builds a transition system to describe the event log, and then group the state regions into corresponding Petri net node. Language-based algorithms uses integer linear system to represent the place constraint where the token at one place can never go negative. By solving the system, a petri net is created. Its representative techniques are Integer Linear Programming(ILP) Miner[19]. Other methods due to [20] include search-based algorithm like Genetic Algorithm Miner[5], heuristic-based algorithm Heuristics Miner[22].

Among those discovery methods, Inductive Miner is widely applied[12]. It investigates the activity order in the traces and represents the order in a directly-follows graph. Based on the graph, it finds the most prominent split from the set of exclusive choice, sequence, parallelism and loop splits on the event log. Afterwards, the corresponding operator to the split is used to build a block-structured process model called process tree. Iteratively, the split sublogs are passed as inputs for the same procedure until single activity is reached and no split is available. A process tree is output as the mined process model.

When the actual event log differs a lot from the referred process model, it is suitable to use the rediscovery method to improve the business execution. However, in some cases, the process enhancement focuses to extend or improve an existing process model by using an actual event log[15]. Besides extending the model with more data perspectives, repair is another type of enhancement. It modifies the model to reflect observed behavior while keeping the model as similar as possible to the original model.

In [8], model repair is firstly introduced into process enhancement. By using conformance checking, the deviations of the event log and process model are detected. The consecutive deviations in log only are collected in the form of subtraces at specific location Q in the model. Later, the subtraces are grouped into sublog that share the same location Q for subprocess discovery. In the earlier version in [8], the sublogs are obtained in a greedy way, while in [9], sublogs are gathered by using ILP Miner to guarantee the fitness. Additional subprocesses and loops are introduced into the existing model to ensure the

fitness, which also brings variants of execution paths into the model.

Later, compared to [8, 9], where all deviations are incorporated in model repair, [6] considers the impact of negative information. In [6], the deviations of the model and event log are firstly analyzed, in order to find out which deviations enforces the positive performance. Given a trace and a selected KPI, an observation instance is built to correlate the number of each log move with KPI output. Based on the observation instance, a set of rules are derived in the form of a decision tree. According to the rules, the original event log is divided into sublogs with traces matching the rules. The sublogs are then repaired to contain only trace deviations which have a positive KPI output. Following repair, the sublogs are merged as the input for model repair in [9]. According to the study case in [6], it provides better result than [9] on the aspect of performance.

As described above, the state-of-the-art repair techniques are based on positive instances, meanwhile the negative information are neglected. Without negative information, it is difficult to balance the fitness and precision of those model. Likewise, few researches give a try to incorporate negative information in multiple forms on process discovery.

In [11], the negative information is artificially generated by analyzing the available events set before and after one position and represented in the form of the complement of positive event sets. Based on the positive and negative event sets, Inductive Logic Programming is applied to detect the preconditions for each activity. Those preconditions are then converted to Petri net after applying a pruning and post-process step. Similar work on model discovery based on artificial negative events are published later. In [21], the author improves the method in [11] by assigning weights on artificial events with respect to unmatching window, in order to offer generalization on model.

The work in [13] uses traces in the event log with negative outcomes as negative information. It extends the techniques of numerical abstract domains and Satisfiability Modulo Theories(SMT) proposed in [3] to incorporate negative information for model discovery. Each trace as positive or negative is transformed as one point in n-dimensional space, n is the number of distinct activities. The execution of a trace reflects the token transmission and marking limits on places in the model. Those limits are represented into the a set of marking inequalities and in a form of convex polyhedron in n-dimensional space. Given half-space hypotheses, SMT solves the inequalities and gives the limits on the process model. Before SMT, negative information is incorporated to shift and rotate the polyhedron, which limits the generalization of the solution space. Because half-space is used, this method can not deal with negative instances overlapped into positive instances.

However, the field of model repair which considers the negative information is new. Furthermore, the idea to incorporate negative instances on trace level into model repair is innovative.

Chapter 3

Preliminary

This chapter introduces the most important ground concepts and notations that are used in the remainder of this article. Firstly, the data and process models in process mining are described; Later, one related discovery technique is introduced.

3.1 Event Log

Business process in organizations is reflected by its execution of a set of activities. The historical execution data is usually in a form of event logs in information systems and can be used by Process Mining to analyze the business execution. To specify the event log, we begin with formalizing the various notations[16] .

Definition 3.1 (Event)**.** An event corresponds to an activity in business execution and can be marked by e. An event is characterized by attributes, like a timestamp, name, and associated costs, etc. The finite set of events in the process is written as $\mathcal{E}$.

Definition 3.2 (Trace)**.** A trace is a finite sequence of events $\sigma \in \mathcal{E}^*$ with conditions that *(i)* each event can not appear twice in a trace, $\forall i, j, 1 \leq i, j \leq |\sigma|, if\ i \neq j, then\ \sigma(i) \neq \sigma(j)$. *(ii)* one event can only appear in one trace, $\forall e \in \sigma, if\ e \in \sigma\prime, then\ \sigma = \sigma\prime$. A trace also has a set of attributes, which describes the trace characters, like the identifier, the trace cost.

Definition 3.3 (Event Log)**.** An event log L is a set of traces. If an event log contains timestamps, then the ordering in a trace should respect these timestamps.

3.2 Process Models

After gathering the event log from information systems, process mining can discover a process model based on the event log, aims to improve understanding and insights on the business process. Multiple process modeling languages are applied in process mining in the last years, such as Petri net, BPMN models, etc. Petri nets are bipartite graph to describe concurrent systems. BPMN models, fully named Business Process Model and Notation models, include many different elements and are flexible but complex tool to visualize business process. Process tree is based on a tree structure to organize the event relation. In this article, due to simplicity, we focus on Petri net and process tree.

3.2.1 Petri Net

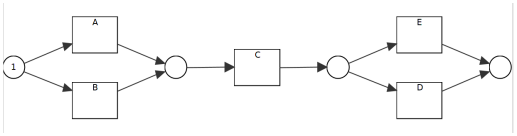
Among multiple models which are proposed to describe business process, Petri net has been best studied thoroughly to allow for the modeling of concurrency of activities and is one of the main process model in process mining. Activities from event log correspond to observable transitions in Petri net. Places in Petri net connect the transitions to express execution rules of activities. Except the observable transitions, there is silent transitions which is not visible.

Definition 3.4 (Silent transition)**.** A silent transition is an internal transition that is non-observable from the outside, but changes the state of the Petri net. It is usually denoted by ϵ.

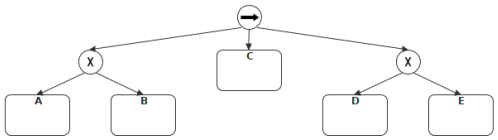
The observable transitions with silent transition ground a basic element set for the static structure of Petri net. To describe the dynamic behaviors of Petri net, the concept called *token* is introduced. It is a mark in the place to enable execution of the following transition. If all the input places for the transition hold a token, the transition is enable, namely the corresponding activity can be triggered. After this execution, the token in the input places are consumed and new tokens are generated in the output places. Initially, only the start place contains a token.

Definition 3.5 (Petri net)**.** A Petri net N is composed of a finite set of places P, transitions T, and a set of directed arcs $F \subseteq (P \times T) \cup (T \times P)$, which can be written as $N = (P, T, F)$. A marked Petri net is (P,T,F,M) where M the marking of the net. A marking of a net N is a multi-set over P, $M \in \mathbb{P}$ and used to express the dynamic state of the Petri net.

An example is shown in Figure 3.1a. It has transitions $T = \{A, B, C, D, E\}$ and four places with the initial marking in the place before $T = \{A, B\}$.



(a) A Simple Petri net



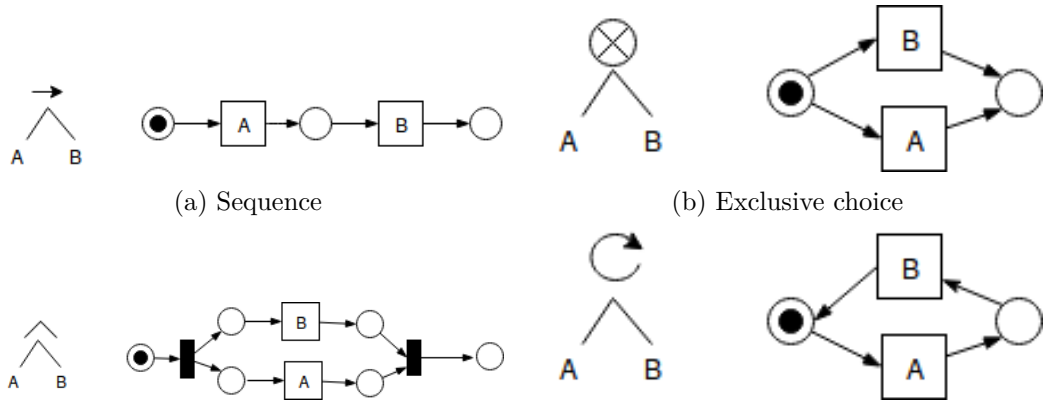
(b) A process tree corresponding to Fig 3.1a

3.2.2 Process Tree

Process tree is block-structured and sound by construction, while Petri nets, BPMN models possibly suffer from deadlocks, other anomalies[16]. Here we give the definition of process tree.

Definition 3.6 (Process Tree)**.** Let $A \subseteq \mathbb{A}$ be a finite set of activities with silent transition $\tau \in \mathbb{A}$, $\bigoplus \subseteq \{\rightarrow, \times, \wedge, \circlearrowleft\}$ be the set of process tree operators.

- $Q = a$ is a process tree with $a \in A$, and
- $Q = \oplus(Q_1, Q_2, ..Q_n)$ is a process tree with $\oplus \in \bigoplus$, and Q_i is a process tree, $i \in 1, 2, .., n, n \in \mathbb{N}$.



(a) Sequence (b) Exclusive choice

(c) Parallelism (d) Loop

Figure 3.2: Semantics of process tree operators w.r.t. Petri net

Process tree operators represents different block relation of each subtree. Their semantics are standardized from [17, 2] and explained with use of Petri net in Figure 3.2[2].

Definition 3.7 (Operator Semantics)**.** The semantics of operators $\bigoplus \subseteq \{\rightarrow, \times, \wedge, \circlearrowleft\}$ are,

- if $Q = \rightarrow (Q_1, Q_2, ..Q_n)$, the subtrees have sequential relation and are executed in order of $Q_1, Q_2, ..Q_n$
- if $Q = \times(Q_1, Q_2, ..Q_n)$, the subtrees have exclusive choice relation and only one subtree of $Q_1, Q_2, ..Q_n$ can be executed.
- if $Q = \wedge(Q_1, Q_2, ..Q_n)$, the subtrees have parallel relation and $Q_1, Q_2, ..Q_n$ they can be executed in parallel.
- if $Q = \circlearrowleft (Q_1, Q_2, ..Q_n)$, the subtrees have loop relation and $Q_1, Q_2, ..Q_n$ *with* $n \geq 2$, Q_1 is the do-part and is executed at least once, $Q_2, ..Q_n$ are redo part and have exclusive relation.

According to the corresponding semantic relations, a process tree can be easily transformed into Petri net. In Figure 3.1b, it is the process model in process tree which describes the same process as in Figure 3.1a.

3.3 Inductive Miner

To discover a process model from an event log, we choose one of the leading process discovery approaches – Inductive Miner, because it guarantees the construction of sound model, and is flexible and scalable to event log data. Its steps are listed bellow.

3.3.1 Construct a directly-follows graph

A the start, the event log L is scanned to extract the directly follows relation of events. The directly-follows relation is like the one in α-algorithm [18, 12], but the frequency information is stored for each relation. Later, those relations are combined together to build a directly-follows graph with frequency. According to [16, 12], a directly-follows graph is defined bellow.

Definition 3.8 (Directly-follows Graph)**.** The directly-follows relation $a > b$ is satisfied iff there is a trace σ *where*, $\sigma(i) = a$ *and* $\sigma(i+1) = b$. A directly-follows graph of an event log L is $G(L) = (A, F, A_{start}, A_{end})$ where A is the set of activities in L, $F = (a, b) \in A \times A | a >_L b$ is the directly-follows relation, A_{start}, A_{end} are the set of start and end activities respectively.

The frequency information of the directly-follows relation is called cardinality and defined below.

Definition 3.9 (Cardinality in directly-follows graph)**.** Given a directly-follows graph G(L) derived from an event log L, the cardinality of each directly-follows relation in G(L) is :

- $Cardinality(E(A, B))$ is the frequency of traces with $\langle ..., A, B, ... \rangle$.
- Start node A cardinality $Cardinality(Start(A))$ is the frequency of traces with begin node A.
- End node B cardinality $Cardinality(End(A))$ is the frequency of traces with end node B.

3.3.2 Split Log Into Sublogs

Based on the directly-follows graph, it finds the most prominent cut which is applied afterwards to split the event log into smaller sublogs. Cuts compose of *exclusive-choice cut, sequence cut, parallel cut and redo-loop cut* which correspond to the process tree operators $\{\rightarrow, \times, \wedge, \circlearrowleft\}$. They are selected in the following order. A maximal exclusive-choice cut is firstly tried to split the directly-follows graph; if it is not available, then a maximal sequence cut, a maximal parallel cut and a redo-loop cut are applied in sequence. Sublogs are created due to this available operator. Meanwhile, this operator is used to build the process tree.

The same procedure is applied again on the sublogs until single activities. What's more, this process tree can be converted into Petri net for further analysis.

3.4 Model Measurements

After deriving a process model from an event log, multiple measurements are used to address the different aspects of the process model quality.

3.4.1 Soundness

Soundness is used to describe if a Petri net is a workflow net with unique start place and unique end place. To prove the model sound, we need to prove the four conditions.

- safeness. Places cannot hold multiple tokens at the same time
- proper completion. If the sink place is marked, all other places are empty.
- option to complete. It is always possible to reach the final marking just for the sink place/
- no dead part. For any transition there is a path from source to sink place through it.

3.4.2 Fitness

3.4.3 Confusion Matrix

Confusion matrix is used from ??

Chapter 4

Algorithm

This chapter describes the repair algorithm to incorporate the negative instances on process enhancement. At the beginning, the main architecture is listed to provide an overview of our strategy. Main modules of the algorithm are described in the next sections. Firstly, the impact of the existing model, positive and negative instances are balanced in the media of the directly-follow relations. Inductive Miner is then applied to mine process models from those directly-follows relations. Again, we review the negative instances and express its impact by adding the long-term dependency. To add long-term dependency, extra places and silent transitions are created on the model, aiming to enforce the positive instances and block negative instances. Furthermore, the model in Petri net with long-term dependency can be post-processed by reducing the silent transitions for the sake of simplicity.

4.1 Architecture

Figure 4.1 shows the steps of our strategy to enhance a process model. The basic inputs are an event log, and a Petri net. The traces in event log have an attribute for the classification labels of positive or negative in respect to some KPIs of business processes. The Petri net is the referenced model for the business process. To repair model with negative instances, the main steps are conducted.

- *Generate directly-follows graph* Three directly-follows graph are generated respectively for the existing model, positive instance and negative instances from event log.

- *Repair directly-follows graph* The three directly-follows graphs are combined into one single directly-follows graph after balancing their impact.

- *Mine models from directly-follows graph* Process models are mined by Inductive Miner as intermediate results.

- *Add long-term dependency* Long-term dependency is detected on the intermediate models and finally added on the Petri net. To simplify the model, the reduction of silent transitions can be applied at end.

More details can be provided in the following sections.

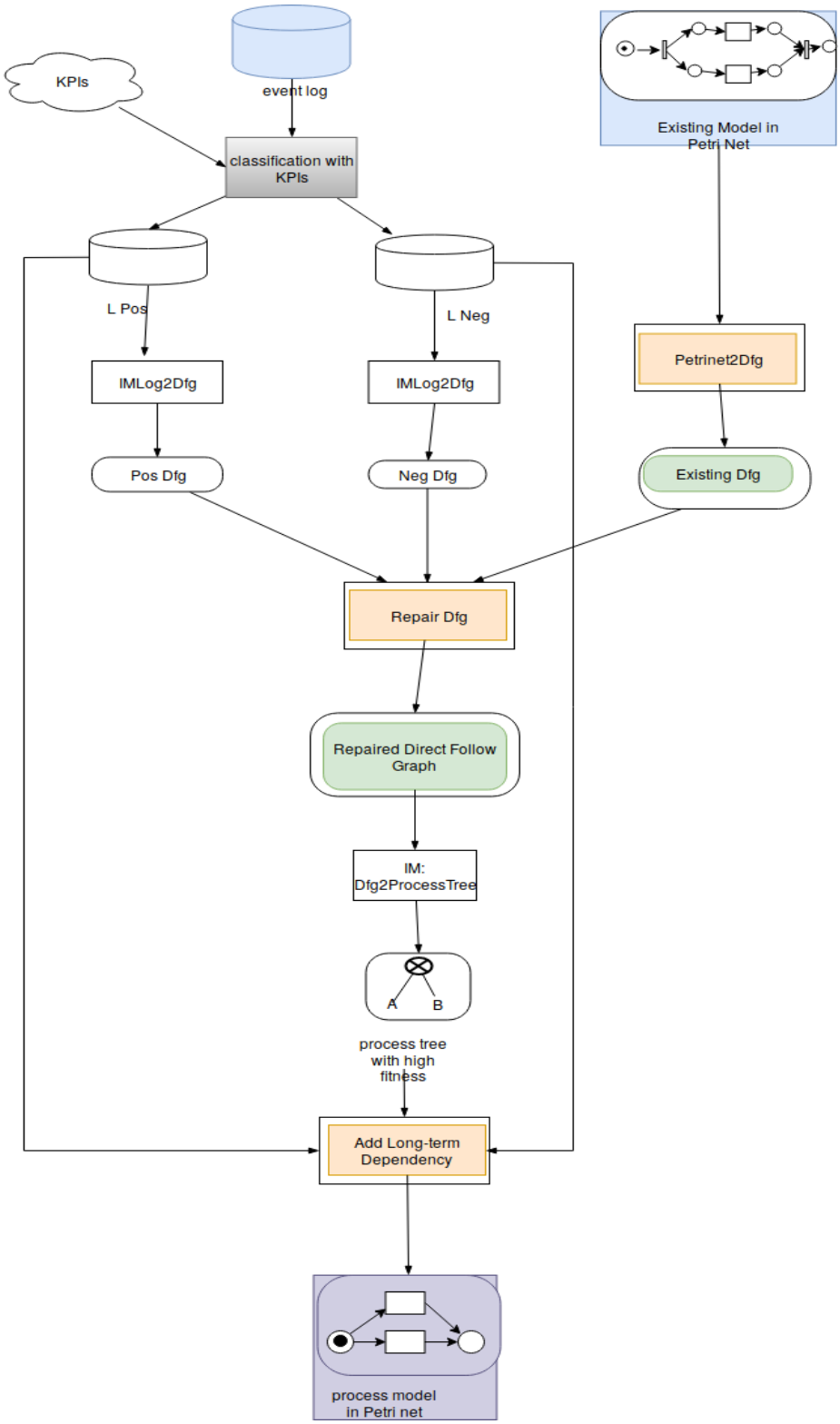


Figure 4.1: Model Repair Architecture – Rectangles represents processes and output data in eclipse shape, especially customized processes and data are in doubled lattice shape. Input event log and existing model are in blue, KPIs are in cloud. The output is a petri net in purple.

4.2 Generate directly-follows graph

Originally, the even log L is split into two sublogs, called L_{pos} and L_{neg}. L_{pos} contains the traces which is labeled as positive, while L_{neg} contains the negative instances in the event log. Then, the two sublogs are passed to procedure *IMLog2Dfg* to generate directly-follows graphs, respectively $G(L_{pos})$ and $G(L_{neg})$. More details about the procedure is available in [12].

To generate a directly-follows relation from a Petri net, we gather the model behaviors by building a transitions system of its states. Then the directly-follows relations are extracted from state transitions. Based on those relations, we create a directly-follows graph for the existing model.

From the positive and negative event log, we can get the cardinality for corresponding directly-follows graph, to represent the strength of this directly-follows relation. However, when the existing model is transformed into directly-follows graph $G(L_{ext})$, there is no point to assign cardinality on each edge. So we just set cardinality with 1 for each edge.

4.3 Repair directly-follows graph

To combine all information of the directly-follows graphs from the positive , negative instances and the existing model, namely $G(L_{pos})$, $G(L_{neg})$ and $G(L_{ext})$, the cardinality in directly-follows graphs is unified into the same range [0-1]. Since $G(L_{ext})$ is derived differently, its unified cardinality is based only on the given directly-follows graph and defined in the following.

Definition 4.1 (Unified cardinality for the existing Model)**.** Given a directly-follows graphs $G(L_{ext})$ for a model, the unified cardinality of each directly-follows relation is defined as

$$U_{ext}(E(A,B)) = \frac{Cardinality(E(A,B))}{Cardinality(E(A,*))}, with$$

$$Cardinality(E(A,*)) = \sum Cardinality(E(A,X)|E(A,X) \in G(L))$$

for start activities A,

$$U(Start(A)) = \frac{Cardinality(Start(A))}{Cardinality(Start(*))}$$

Similarly for end activities B,

$$U(End(B)) = \frac{Cardinality(End(B))}{Cardinality(End(*))}$$

$E(A,*)$ means all edges with source A, $E(*,B)$ means all edges with target B, $Start(*)$ represents all start nodes, and $End(*)$ represents all end nodes.

The unification of cardinality for positive and negative instances from an event log should consider the whole event log as its basic.

Definition 4.2 (Unified cardinality for $G(L_{pos})$, $G(L_{neg})$)**.** Given a directly-follows graphs $G(L_{pos})$ for a model, the unified cardinality of each directly-follows relation is defined as

$$U_{pos}(E(A,B)) = \frac{Cardinality_{pos}(E(A,B))}{Cardinality(E(A,*))}, with$$
$$Cardinality(E(A,*)) = \sum Cardinality_{pos}(E(A,X) + Cardinality_{neg}(E(A,Y)) where$$
$$E(A,X) \in G(L_{pos}) and E(A,Y) \in G(L_{neg}))$$

for start activities A,

$$U(Start_{pos}(A)) = \frac{Cardinality_{pos}(Start(A))}{Cardinality(Start(*))}$$

for end activities B,

$$U(End_{pos}(B)) = \frac{Cardinality_{pos}(End(B))}{Cardinality(End(*))}$$

The unification for negative instances is defined in a similar way.

Considering all the unified cardinalities, we derive a concept called weight for directly-follows relation to combine the factors from the existing model, positive and negative instances. Later, a new directly-follows graph is built based on those weights.

Definition 4.3 (Weight of directly-follows relation G_{new})**.**

- For one directly-follows relation,

$$Weight(E_{G_{new}}(A,B)) = U(E_{G_{ext}}(A,B)) + U(E_{G_{pos}}(A,B)) - U(E_{G_{neg}}(A,B))$$

- For start acivities A, we have

$$Weight(Start_{G_{new}}(A)) = U(Start_{G_{ext}}(A)) + U(Start_{G_{pos}}(A)) - U(Start_{G_{neg}}(A))$$

- For end activities B, we have

$$Weight(End_{G_{new}}(A)) = U(End_{G_{ext}}(A)) + U(End_{G_{pos}}(A)) - U(End_{G_{neg}}(A))$$

In the real life, there exists various needs to address either on the existing model, the positive instances or the negative instances. To meet this requirement, three control parameters are assigned respectively to each unified cardinality from the existing model, and positive and negative instances. The weight for one directly-follow relation is modified in the way bellow.

Definition 4.4 (Weight of directly-follows relation G_{new})**.**

- For one directly-follows relation,

$$Weight(E_{G_{new}}(A,B)) = C_{ext}{*}U(E_{G_{ext}}(A,B)) + C_{pos}{*}U(E_{G_{pos}}(A,B)) - C_{neg}{*}U(E_{G_{neg}}(A,B))$$

- For start acivities A, we have

$$Weight(Start_{G_{new}}(A)) = C_{ext}{*}U(Start_{G_{ext}}(A)) + C_{pos}{*}U(Start_{G_{pos}}(A)) - C_{neg}{*}U(Start_{G_{neg}}(A))$$

- For end activities B, we have

$$Weight(End_{G_{new}}(A)) = C_{ext}{*}U(End_{G_{ext}}(A)) + C_{pos}{*}U(End_{G_{pos}}(A)) - C_{neg}{*}U(End_{G_{neg}}(A))$$

By adjusting the weight of C_{ext},C_{pos},C_{neg}, different focus can be reflected by the model. For example, by setting $C_{ext} = 0$,$C_{pos} = 1$,$C_{neg} = 1$, the existing model is ignored in the repair, while $C_{ext} = 1$,$C_{pos} = 0$,$C_{neg} = 0$, the original model is kept.

4.4 Mine models from directly-follows graph

The result from last step is a generated directly-follows graph with weighted unified cardinality. To apply the Inductive Miner on directly-follows graph, more procedures are in need. The directly-follows relations are firstly filtered when its weighted unified cardinality is less than 0. Secondly, its cardinality is transformed into the form which is acceptable by the Inductive Miner.

4.5 Add long-term dependency

Due to the intrinsic characters of Inductive Miner, the dependency from activities which are not directly-followed can not be discovered. To make the generated model more precise, we detect the long-term dependency and add it on the model in Petri net. Obviously, long-term dependency relates the choices structure in process model, such as exclusive choice, loop and or structure. Due to the complexity of or and loop structure, the long-term dependency in exclusive choice is considered only in this thesis.

To analyze the exclusive choice of activities, we uses process tree as a intermediate process. We use process trees as one internal result in our approach in two factors: The reasons are: (1) easy to extract the exclusive choice structure from process tree. Process tree is block-structured and benefits the analysis of exclusive choices. (2) easy to get the model in process tree. Inductive Miner can generate process model in process tree. (3) easy to transform process tree to Petri net. The exclusive choice can be written as Xor in process tree. For sake of convenience, we define the concept called xor branch.

Definition 4.5. Xor branch $Q = \times(Q_1, Q_2, ..Q_n)$, Q_i is one xor branch with respect to Q, rewritten as $XORB_{Q_i}$ to represent one xor branch Q_i in xor block, and record it $XORB_{Q_i} \in XOR_Q$. For each branch, there exists the begin and end nodes to represent the beginning and end execution of this branch, which is written respectively as Begin($XORB_{Q_i}$) and End($XORB_{Q_i}$).

Two properties of xor block, purity and nestedness are demonstrated to express the different structures of xor block according to its branches.

Definition 4.6 (XOR Purity and XOR Nestedness)**.** The xor block purity and nestedness are defined as following:

- A xor block XOR_Q is pure if and only $\forall XORB_X \in XOR_Q, XORB_X$ has no xor block descent, which is named as pure xor branch. Else,
- A xor block XOR_Q is nested if $\exists XOR_X, Anc(XOR_Q, XOR_X) \rightarrow True$. Similarly, this xor branch with xor block is nested.

For two arbitrary xor branches, to have long-term dependency, they firstly need to satisfy the conditions: (1) they have an order;(2) they have significant correlation. The order of xor branch follows the same rule of node in process tree which is explained in the following.

Definition 4.7 (Order of nodes in process tree)**.** Node X is before node Y, written in $X \prec Y$, if X is always executed before Y. In the aspect of process tree structure, $X \prec Y$, if the least common ancestor of X and Y is a sequential node, and X positions before Y.

The correlation of xor branches is significant if they always happen together. To define it, several concepts are listed at first.

Definition 4.8 (Xor branch frequency)**.** Xor branch $XORB_X$ frequency in event log L is $F_L(XORB_X)$, the count of traces with the execution of $XORB_X$.
For multiple xor branches, the frequency of their coexistence in event log L is defined as the count of traces with all the occurrence of xor branches $XORB_{Xi}$, written as

$$F_L(XORB_{X1}, XORB_{X2}, ..., XORB_{Xn})$$

.

After calculation of the frequency of the coexistence of multiple xor branches in positive and negative event log, we get the supported connection of those xor branches to reflect the correlation.

Definition 4.9 (Correlation of xor branches)**.** For two pure xor branches, $XORB_X \prec XORB_Y$, the supported connection is given as

$$SC(XORB_X, XORB_Y) = F_{pos}(XORB_X, XORB_Y) - F_{neg}(XORB_X, XORB_Y)$$

If $SC(XORB_X, XORB_Y) >$ lt-threshold, then we say $XORB_X$ and $XORB_Y$ have significant correlation.

The existing model can also affect the long-term dependency by supporting the existence of xor branches. In this assumption, the full long-tern dependency is approved. To incorporate the influence from the existing model, we rephrase the definition for xor branch correlation.

Definition 4.10 (Rephrased Correlation of xor branch)**.** The correlation for two branches is expressed into

$$Wlt(XORB_X, XORB_Y) = Wlt_{ext}(XORB_X, XORB_Y) + Wlt_{pos}(XORB_X, XORB_Y)$$

$$-Wlt_{neg}(XORB_X, XORB_Y)$$

, where $Wlt_{ext}(XORB_X, XORB_Y) = \frac{1}{|XORB_{Y*}|}$, $|XORB_{Y*}|$ means the number of possible directly-follows xor branche set $XORB_{Y*} = \{XORB_{Y1}, XORB_{Y2}, ...XORB_{Yn}\}$ after $XORB_X$.
$Wlt_{pos}(XORB_X, XORB_Y) = \frac{F_{pos}(XORB_X, XORB_Y)}{F_{pos}(XORB_X, *)}$,
$Wlt_{neg}(XORB_X, XORB_Y) = \frac{F_{neg}(XORB_X, XORB_Y)}{F_{neg}(XORB_X, *)}$,

The $F_{pos}(XORB_X, XORB_Y)$ and $F_{neg}(XORB_X, XORB_Y)$ are the frequency of the coexistence of $XORB_X$ and $XORB_Y$, respectively in positive and negative event log.

4.5.1 Cases Analysis

There are 7 sorts of long-term dependency that is able to happen in this model as listed in the following. Before this, we need to define some concepts at the sake of convenience.

Definition 4.11 (Source and Target of Long-term Dependency)**.** We define the source set of long-term dependency is $LT_S := \{X | \exists X, X \rightsquigarrow Y \in LT\}$, and target set is $LT_T := \{Y | \exists Y, X \rightsquigarrow Y \in LT\}$.
For one xor branch $X \in XORB_S$, the target xor branch set relative to it with long-term dependency is defined as: $LT_T(X) = \{Y | \exists Y, X \rightsquigarrow Y \in LT\}$ Respectively, the source xor branch relative to one xor branch in target is $LT_S(Y) = \{X | \exists X, X \rightsquigarrow Y \in LT\}$

At the same time, we use $XORB_S$ and $XORB_T$ to represent the set of xor branches for source and target xor block.

1. $LT = \{A \rightsquigarrow D, A \rightsquigarrow E, B \rightsquigarrow D, B \rightsquigarrow D\}$.
   $LT_S = \{A, B\}, LT_T = \{D, E\}, |LT| = |XORB_S| * |XORB_T|$, which means long-term dependency has all combinations of source and target xor branches.

2. $LT = \{A \rightsquigarrow D, A \rightsquigarrow E, B \rightsquigarrow E\}$.
   $LT_S = \{A, B\}, LT_T = \{D, E\}$ $LT_S = XORB_S$ and $LT_T = XORB_T, |LT| < |XORB_S| * |XORB_T|$. it doesn't cover all combinations. But for one xor branch $X \in XORB_S, LT_T(X) = XORB_T$, it has all the full long-term dependency with $XORB_T$.

3. $LT = \{A \rightsquigarrow D, B \rightsquigarrow E\}$.
   $LT_S = \{A, B\}, LT_T = \{D, E\}$ $LT_S = XORB_S$ and $LT_T = XORB_T, |LT| < |XORB_S| * |XORB_T|$. For all xor branch $X \in XORB_S, LT_T(X) \subsetneq XORB_T$, none of xor branch X has long-term dependency with $XORB_T$.

4. $LT = \{A \rightsquigarrow D, B \rightsquigarrow D\}$.
   $LT_S = XORB_S, LT_T \subsetneq XORB_T$. There exists at least one xor branch $Y \in XORB_T$ which has no long-term dependency on it.

5. $LT = \{A \rightsquigarrow D, A \rightsquigarrow E\}$.
   $LT_S \subsetneq XORB_S, LT_T = XORB_T$. There exists at least one xor branch in source $X \in XORB_S$ which has no long-term dependency on it.

6. $LT = \{A \rightsquigarrow E\}$.
   $LT_S \subsetneq XORB_S, LT_T \subsetneq XORB_T$. There exists at least one xor branch in source $X \in XORB_S$ and one xor target xor branch which has no long-term dependency on it.

7. $\emptyset$. There is no long-term dependency on this set.

In the following, we propose a method to express long-term dependency on Petri net.

4.5.2 Way to express long-term dependency

In dynamic aspect of the process model, long-term dependency can be seen as a way to block certain behaviors. By injecting silent transitions and extra places on Petri net, it limits the executions of certain behaviors. The steps to add silent transitions and places according to the long-term dependency are listed bellow.

Algorithm 1: Add long-term dependency between pure xor branch

$XORB_Y$ is dependent on $XORB_X$;
if $XORB_X$ *is leaf node* **then**
    One place is added after this leaf node. ;
end
if $XORB_X$ *is Seq* **then**
    Add a place after the end node of this branch;;
    The node points to the new place;;
end
if $XORB_X$ *is And* **then**
    Create a place after the end node of every children branch in this And xor branch; ;
    Combine all the places by a silent transition after those places; ;
    Create a new place directly after silent transition to represent the And xor branch; ;
end
if $XORB_Y$ *is leaf node* **then**
    One place is added before this leaf node. ;
end
if $XORB_Y$ *is Seq* **then**
    Add a place before the end node of this branch;;
    The new place points to this end node;;
end
if $XORB_Y$ *is And* **then**
    Create a place before the end node of every children branch in this And xor branch; ;
    Combine all the places by a silent transition before those places; ;
    Create a new place directly before silent transition to represent the And xor branch; ;
end
Connect the places which represent the $XORB_X$ and $XORB_Y$ by creating a silent transition.

4.5.3 Soundness Analysis

With this algorithm by adding silent transitions and places to express long-term dependency, the model soundness can be violated. In the following section, we will discuss the soundness in different situations.

Given arbitrary two xor block, $S = \{X_1, X2, ...X_m\}$ and $T = \{Y_1, Y_2, ...Yn\}$ with long-term dependency $LT = \{X_i \rightsquigarrow Y_j | 1 \leq i \leq m, 1 \leq j \leq n\}$, following the algorithm1, places after the source xor branches, $P_S = p_{X_i} | X_i \in LT_S$, and places before target xor branches,$P_T = p_{Y_j} | Y_i \in LT_T$ are firstly added.
For each long-term dependency $X_i \rightsquigarrow Y_j$ in LT, there is silent transition t with $p_{X_i} \rightarrow t \rightarrow p_{Y_j}$.

To prove the model soundness after adding long-term dependency, we need to prove

- the algorithm keep soundness between xor blocks.

Proof. For all xor branches in S, only one branch can be enabled. Without loss if generality, X_i is assumed to be enabled. After firing X_i, the marking distribution is those xor block are

$$M(p_{X_i}) = 1, \forall p_{X_{i\prime}} \in P_S, i\prime \neq i, M(p_{X_{i\prime}}) = 0$$

The silent transition t with $p_{X_i} \rightarrow t \rightarrow p_{Y_j}$, can pass one token from p_{X_i} to p_{Y_j}. After firing t, the marking distribution changes to

$$M(p_{Y_j}) = 1, \forall p_{Y_{j\prime}} \in P_T, j\prime \neq j, M(p_{Y_i}) = 0$$

For each xor branch in T, it has two input places, one of them is the new added place $p_{Y_j}, Y_j \in T$. Now only Y_j is enabled. After executing the Y_j, it consumes the extra token at place p_{Y_j}. □

- In Situation 1,2 and 3, $LT_S = S, LT_T = T$, the original model is sound, now only consider the extra places and transitions, if they satisfy the conditions, too, then the whole model is sound.
  - safeness.
    For all extra places p_{X_i} and p_{Y_j},

    $$\forall p_{X_i} \in P_S, \sum M(p_{X_i}) = 1, p_{Y_j} \in P_T, \sum M(p_{Y_j}) = 1$$

  - proper completion.
    After firing Y_j, all the extra places hold no token. So it does not violate the proper completion.
  - option to complete.
    There is always one Y_j enabled to continue the subsequent execution.
  - no dead part.
    Because all $Y_j \in T$ are also in LT_T, there exists at least one $X_i \in S$ with long-term dependency with Y_j. After X_i is fired, one token is generated on the extra place p_{X_i} and can be consumed by silent transition t in $p_{X_i} \rightarrow t \rightarrow p_{Y_j}$ to produce a token in p_{Y_j}, which enables xor branch Y_j and leaves no dead part.
- In the rest situations, $LT_S \neq S = X_i, LT_T \neq \emptyset$, there exists one xor branch$X_i with X_i \notin LT_S, Y_j \in LT_T$, when X_i is selected, it generates one token at place p_{X_i}, this token can not be consumed by any Y_j. So it violates the proper completion. $If \quad LT_T \neq T = Y_j$, there exists one $Y_j \notin LT_T, \nexists X_i, X_i \rightsquigarrow Y_j$, so with two input places but $Token(p_{Y_j}) = 0$, Y_j becomes the dead part, which violates the soundness again.
  In source xor block S, only one xor branch X_i can be triggered.

In the target xor block T, only one xor block Y_j can be triggered.

- the algorithm does not violate the soundness outside of xor blocks.
  the added silent transitions and places do not damage the execution outside of the xor blocks. Because the tokens are only stays in those xor blocks, so it does not affect the execution of other parts. The tokens flow through the network like it does before.

An example is given in Figure **??**. Given the long-term dependency in situation 2 with $LT = \{A \rightsquigarrow D, A \rightsquigarrow E, B \rightsquigarrow E\}$, firstly, two extra places are added respectively after A and B; Next, two places before D and E are created to express that the xor branches are involved with long-term dependency. At end, for each long-term dependency with xor branches, a silent transition is generated to connect the extra places after the source xor branch to the place before target place.

4.5.4 Feasible cases due to soundness

However, only with data from positive and negative information on the long-term dependency, it is possible to result in unsound model, because some directly-follows relations about xor branches can be kept due to the existing model. When those relations do not show in the positive event log or shows only in negative event log, we get incomplete information and no evidence of long-term dependency is shown on those xor branches. In this way, it results in an unsound model, since those xor branches can't get fired to consume the tokens generated from the choices before.

For situation 1, it's full connected and xor branches can be chosen freely. So there is no need to add explicit connection on model to represent long-term dependency, therefore the model keeps same as original. For Situation 2 and 3, if $LT_S = XORB_S, LT_T = XORB_T$, then we can create an sound model by adding silent transitions. If not, then we need to use the duplicated transitions to create sound model. But before, we need to prove its soundness. For situation 4, 5 and 6, there is no way to prove the soundness even by adding duplicated events.

By adding constraints, we make sure only situations 1,2,3 happen when adding long-term dependency.

4.6 Reduce Silent Transitions

Our method to represent long-term dependency can introduce redundant silent transitions and places. On one hand, it complicates the model; on the other hand, it causes the model unsound where the extra silent transitions suspend the execution and therefore violates the soundness condition of proper completion. So, in this step, we postprocess the Petri net to reduce silent transitions.

Proposition 4.12. Given a silent transition ϵ in Petri net with one input place P_{in} and one output place P_{out}, if $|Outedges(P_{in})| \geq 2 \quad and \quad |Inedges(P_{out})| \geq 2$, the silent transitions can not be deleted. Else, the silent transitions is able to delete, meanwhile the $P_{in} and P_{out}$ can be merged into one place. This reduction does not violate the soundness and does not change the model behavior.

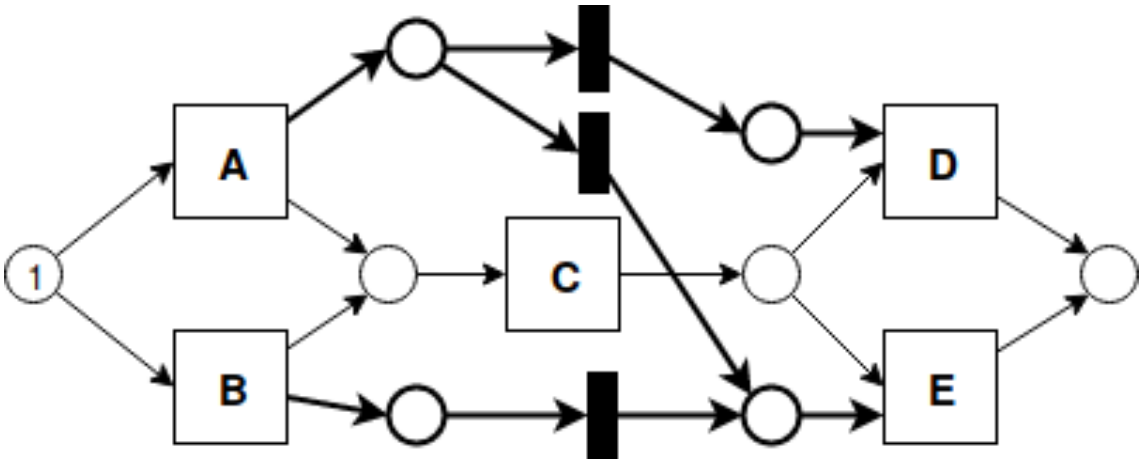
Soundness Proof. If a silent transition t is able to delete, then

$$|Outedges(P_{in})| \leq 1$$

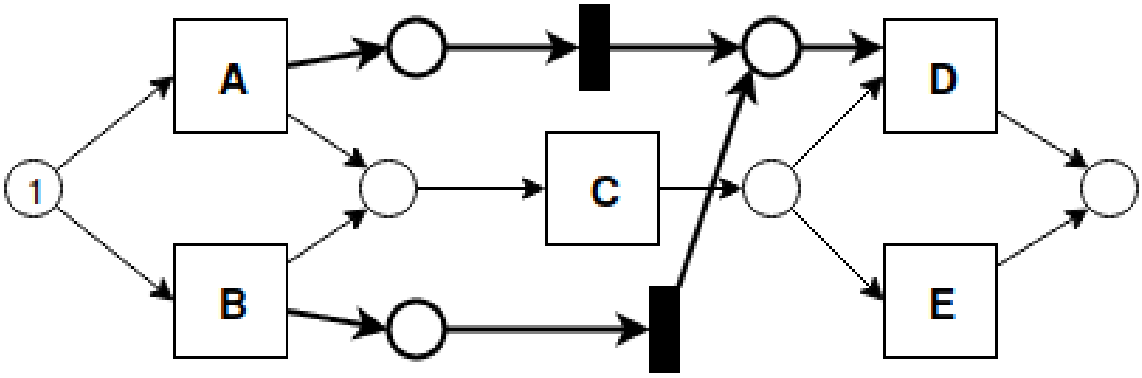
, or

$$|Inedges(P_{out})| \leq 1$$

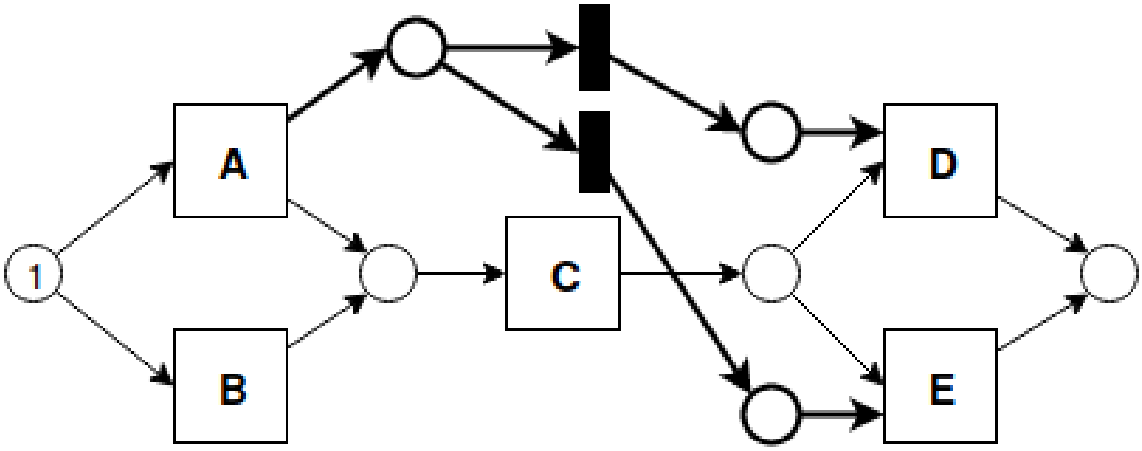
,
when in case (1), P_{in} contains a token that is always passed to P_{out} by silent transitions t. After deleting the silent transition, the token is generated directly on P_{out}. Since t is



(a) For situation 2



(b) For situation 4



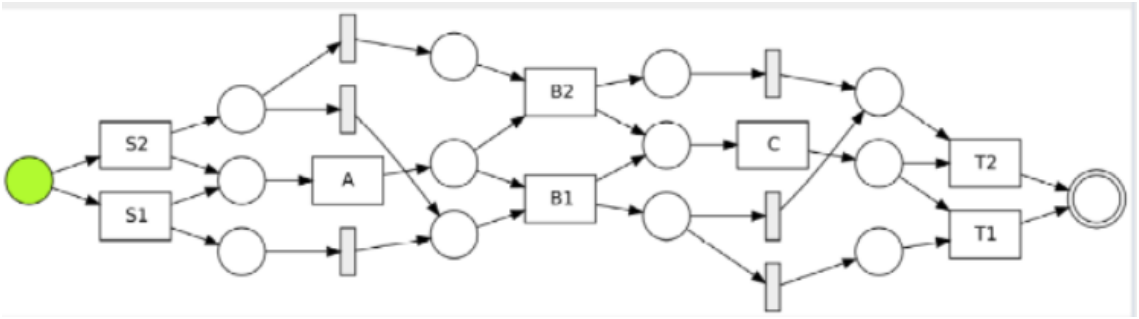
(c) For situation 5

Figure 4.2: Silent Events for Long-term Dependency

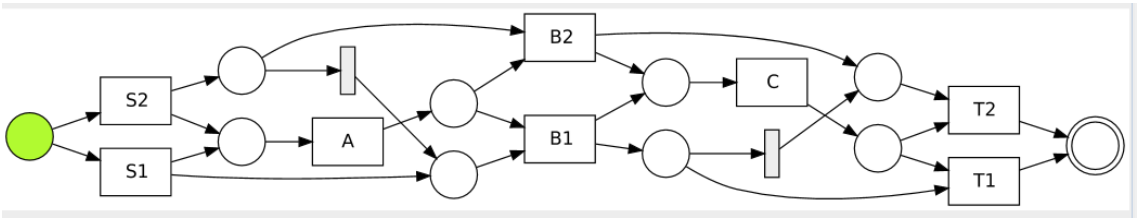
silent transition, it won't affect the model behavior.
when in case (2), P_{out} contains a token that is always passed from P_{in} by silent transitions

t. After deleting the silent transition, the token is remained on P_{in}, which enables the later execution after the original P_{out}. Since t is silent transition, it won't affect the model behavior. □

One example is given in the following graph.



(a) A Petri net with redundant silent transitions



(b) A Petri net with reduced silent transitions

Chapter 5

Implementation

ProM is an open-source extensible framework. It supports wide process mining techniques in the form of plug-ins[14]. The algorithm to incorporate negative information is implemented as one plug-in called ***Repair Model By Kefang*** in ProM and released online[7]. This chapter is divided into four parts to describe the whole implementation. Firstly, the inputs and outputs of this implementation is introduced. After accepting the inputs, a directly-follows graph is constructed by dfg-method and later converted into process tree or Petri net process model. Next, the implementation to add long-term dependency on the process model from last step is shown. At last, another feature is displayed to show the brief evaluation result based on confusion matrix.

5.1 Inputs And Outputs

The plug-in inputs are an event log L with labels to classify each trace and an existing model N possible in multiple forms.

- Acceptable event log L with labels. Labels are one trace attribute and used to identify the positive and negative instances.
- Acceptable Process Models
  - Petri net + Initial Marking.
  - Accepting Petri net.
    The initial marking is already included into the accepting Petri net.
  - Petri net.
    In this situation, the initial marking is guessed automatically in the background program.

The outputs are one process model and its corresponding initial marking. They are exported from the control panel in the result view and can be in various forms.

- Petri net with long-term dependency after Reducing Silent Transition
- Petri net with long-term dependency
- Petri net without long-term dependency
- Process tree

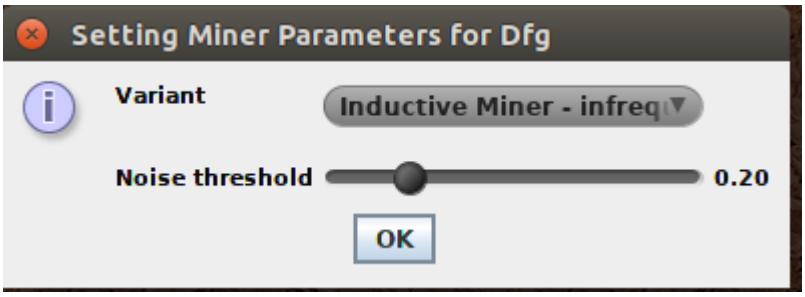


Figure 5.1: Inductive Miner Parameter Setting

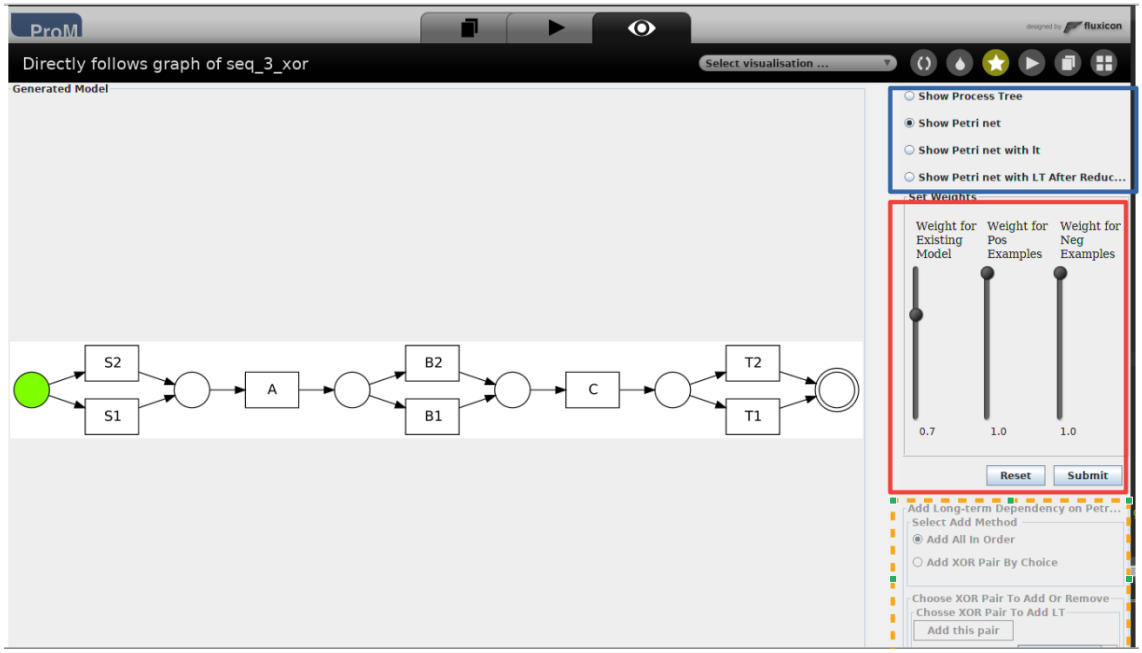


Figure 5.2: Generated Petri net without long-term dependency

5.2 Implementation of dfg-method

Firstly the dialog for options to generate directly-follows graphs from event log pops up. Event classifier are set by those dialogs. Subsequently, a dialog is shown to set the Inductive Miner parameters. The parameters include the Inductive Miner variant and the noise threshold to filter the data. The dialog is displayed in Figure 5.1.
After setting the parameters, process models of process tree and Petri net without long-term dependency can be generated by Inductive Miner and displayed in the result view in Figure 5.2. The left side is the model display area. To allow more flexibility, this plug-in are interactive by the control panel, which is the right side of result view. Originally, only the generated model type and the weight sliders are enabled, while the control panel for adding long-term dependency are invisible.

The model type are in the blue rectangle marked in Figure 5.2. It has 4 options to control the generated model type. Currently, the option "Show Petri net" is chosen, so the constructed model is Petri net without long-term dependency. The weights sliders are in red rectangle. It enables to adjust the weights on the existing model, positive

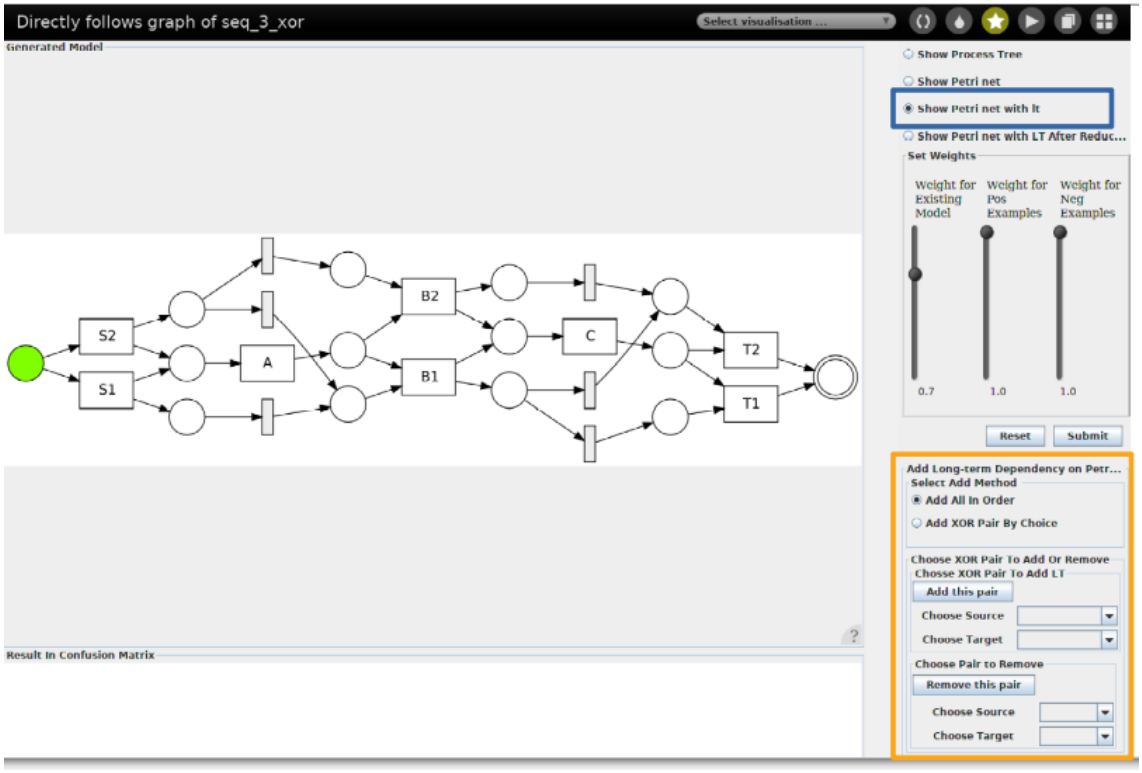


Figure 5.3: Petri Net with long-term dependency

and negative instances. Once submitted those options, different process models are mined under different weights. The rectangle in orange are the invisible part to control long-term dependency options. It is discussed in the next section.

5.3 Implementation of Adding Long-term Dependency

If the model ro generate is Petri net with long-term dependency, the program to add long-term dependency is triggered. This program in the background detects and puts places and silent transitions on Petri net directly mined from Inductive Miner to add long-term dependency. As comparison, the same weight setting is kept like the Figure 5.2, but the option to show a Petri net with long-term dependency is chosen. The resulted model is Figure 5.3.

Meanwhile, the control part of adding long-term dependency turns visible, which is in the orange rectangle in Figure 5.3. It has two main options, one is to consider all long-term dependency existing in the model, the other is to choose the part manually. It allows more flexibility for users. Below those two options, it is the manual selection panels, including control part to add and remove pair. As an example, the blocks Xor(S1,S2) and Xor(T1,T2) are chosen to add long-term dependency. It results in the model in Figure 5.4.

By choosing *Petri net with LT After Reducing* in model type option panel, silent transitions are reduced to simplify the model. Under the same setting in Figure 5.2, the simpler model in Figure 5.5 is constructed, after the post processing of reducing silent transitions.

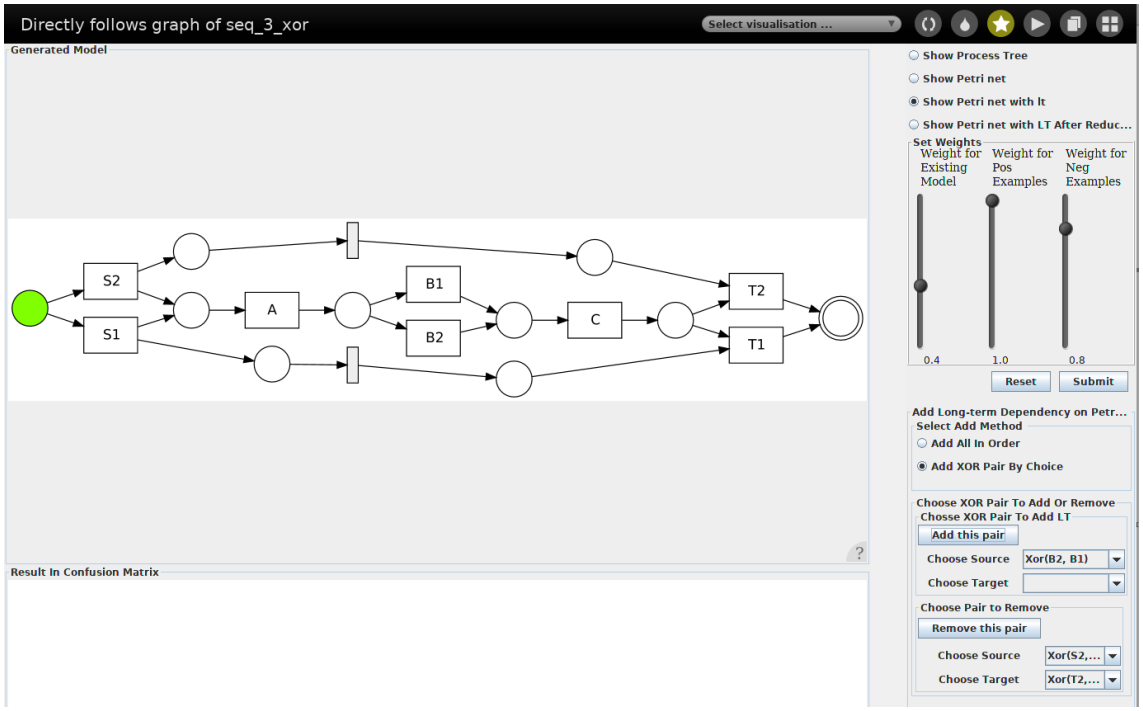


Figure 5.4: Petri net with selected long-term dependency

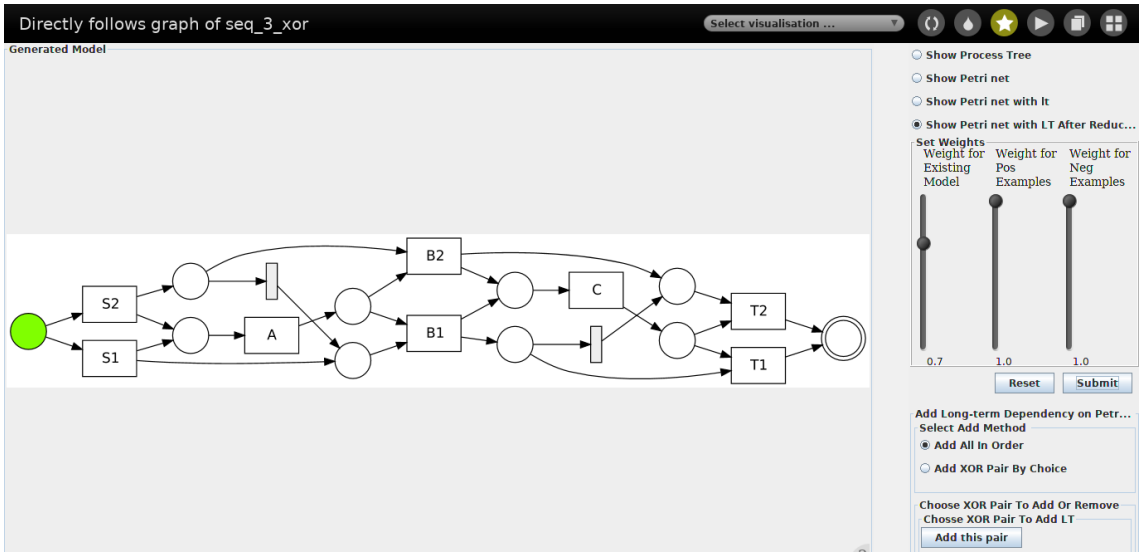


Figure 5.5: Petri net after reducing the silent transitions

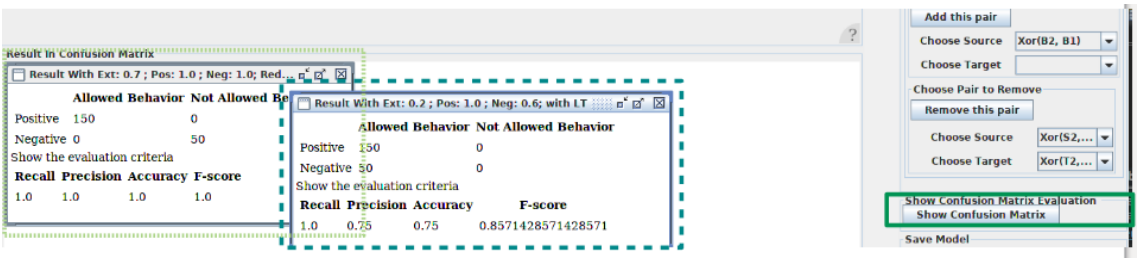


Figure 5.6: Generated Process Tree Model

5.4 Implementation of Showing Evaluation Result

Another feature in this plugin is to show the evaluation result based on confusion matrix. With the brief evaluation result, it helps set the parameter and select the final process model.

It works in this way. After creating the current model in the left view, the evaluation program in background uses the event log and the current Petri net in the view as inputs. It applies a naive fitness checking and generates a confusion matrix with relative measurements like recall, precision. This evaluation result is then shown in the bottom of the left view in Figure 5.6. If the button of green rectangle in the right view *Show Confusion Matrix* is pressed again, the program is triggered again and generates a new confusion matrix result in dark green dashed rectangle which will be listed above the previous result in light green dashes area.

Chapter 6

Evaluation

This chapter presents an experimental evaluation of our techniques to repair model. At first, the evaluation measurements are defined. Next, we briefly introduce the test platform KNIME and ProM plugins tools for evaluation. In the following main part, the test on properties of our techniques is presented at the beginning. Then synthetic data is generated randomly to show the whole performance of our methods. At last, we conduct our experiments on real life data and also compare our techniques with other methods. The results show the ability of our techniques to repair model with high ranking according to defined measurements.

6.1 Evaluation Measurements

We evaluate our techniques based on the quality of repaired models with respect to the given event logs. In process mining, there are four quality dimensions generally used to compare the process models with event logs.

- *fitness.* It quantifies the extent of a model to reproduce the traces recorded in an event log which is used to build the model. Alignment-based fitness computation aligns as many events from trace with the model execution as possible.
- *precision.* It assesses the extent how the discovered model limits the completely unrelated behavior that doesn't show in the event log.
- *generalization.* It addresses the over-fitting problem when a model strictly matches to only seen behavior but is unable to generalize the example behavior seen in the event log.
- *simplicity.* This dimension captures the model complexity. According to Occam's razor principle, the model should be as simple as possible.

The four traditional quality criteria are proposed in semi-positive environment where only positive instances are available. Therefore, when it comes to the model performance, where negative instances are also possible. The measurement metrics should be adjusted. The repair techniques in this thesis are based on the labeled data and the repaired model can be seen as a binary prediction model where the positive instances are supported while the negative ones are rejected. The model evaluation becomes a classification evaluation. Confusion matrix has a long history to evaluate the performance of a classification model.

Table 6.1: Confusion Matrix

		repaired model	
		allowed behavior	not allowed behavior
actual data	positive instance	TP	FN
	negative instance	FP	TN

A confusion matrix is a table with columns to describe the prediction model and rows for actual classification on data. The repaired model can be seen a binary classifier and produces four outcomes- true positive, true negative, false positive and false negative shown in the Table 6.1.

- True Positive(TP): The execution allowed by the process model has an positive performance outcome.
- True Negative(TN): The negative instance is also blocked by the process model.
- False Positive(FP): The execution allowed by the process model has an negative performance outcome.
- False Negative(FN):The negative instance is enabled by the process model.

Various measurements can be derived from confusion matrix. According to our model, we choose the following ones as the potential measurements.

- recall. It represents the true positive rate and is calculated as the number of correct positive predictions divided by the total number of positives.

$$Recall = \frac{TP}{TP + FP}$$

- precision. It describes the ability of the repaired model to produce positive instances.

$$Precision = \frac{TP}{TP + FN}$$

- specificity. In opposite with recall, it measures the true negative rate.

$$Specificity = \frac{TN}{TN + FP}$$

- accuracy. It is the proportion of true result among the total number. It measures in our case how well a model correctly allows the positive instances or disallows the negative instances.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

- F-score is is the harmonic mean of precision and recall.

$$F_1 = \frac{2 * Recall * Precision}{Precision + Recall}$$

After experiments we can see that accuracy gaine more weights than precision.

Generally, there is a trade-off between the quality criteria. So the measurements are only used to compare specific aspects of our techniques.

6.2 Experiment Platform

Data simulation part Try to automatize and standardize experiments, in order to make this experiments more effective, and easier to repeat. So KNIME is selected as one of the test platform to perform tasks. Also, some existing evaluation plugins in ProM are used to evaluate on the other aspects.

6.2.1 PLG

Process Log Generator(PLG) is designed to generate random business processes models and event logs according by setting the complexity parameters. Complexity parameters for model generation in PLG include the maximum number of branches for activities relation parallel, or exclusive choices, maximum depth of nested pattern, the weights on relations, and so on. The parameters to generate an event log are, for example, number of traces, and noise on the workflow or missing rate of trace heads.

In this experiment, process models in Petri net are randomly created by inputting simple, normal and complex parameters settings. Event logs are also simulated according to different complexity level. The setting details is in Table??.

6.2.2 KNIME

KNIME Analytics Platform is open source software in Java to help researcher analyze data by integration of multiple modules for loading, process and transformation and machine learning algorithms. Researcher can achieve their goals by creating visual workflows composed of modules with an intuitive, drag and drop style graphical interface, rather than focusing on any particular application area. The reason to choose KNIME as our experiment platform is that KNIME supports automation of test workflow which is more efficient. However, to conduct experiments, the related modules of our algorithm have to be integrated into KNIME, which requires additional development effort. At end, process mining modules like event logs, basic process models reader and writer, basic logs manipulators, classic discovery algorithms, and model repair are imported into KNIME.

6.2.3 ProM Evaluation Plugins

Here we discuss the plugins to test other aspects of our methods.

6.3 Experiment Result

6.3.1 Test On Property

In this experiment, we aim to answer the questions: 1) How do our techniques incorporate the existing model, positive and negative information to repair model? 2) How do the techniques overcome the shortcomings of current state-of-the-art repair techniques? To answer the first question, we applied the repaired techniques on event logs with different relations of activities, such as sequence, parallel and loop, exclusive choice and investigated the effect from the existing model, positive and negative instances on the repaired model, by adjusting the corresponding weights. For the second question, we describe the situations where our techniques overcome the defects of other repair techniques and explain the reasons behind it.

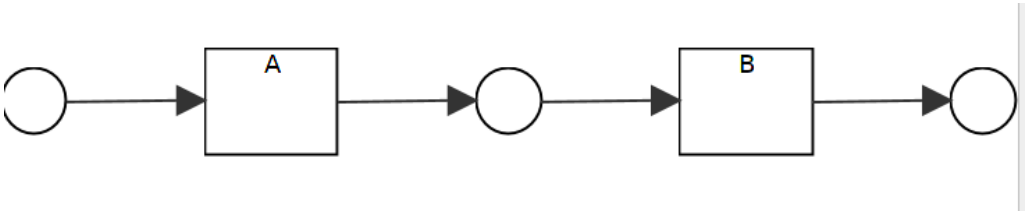


Figure 6.1: Model M1 with sequence relation

Test On Sequence

This part is used to show the effect of our techniques on the sequence relation of activities. Given a fixed model in Petri net with sequence relation, a set of event logs with different deviations are used to test if our repair techniques work properly. By changing the control weights of the existing model, positive and negative instances, different models are derived from the repair method. Next, we evaluate the model based on confusion matrix and plot the several measurements with the weight changes. By this way, we show how our method incorporates different effect of control weights.

In this experiment, we fix one weight as control variable while keeping the other weights unchanged. To make the experiments more reliable, for each fixed control variable value, we conduct the experiment in 5 different combinations of other weights that are randomly generated. At last, the average of those combination is used to represent the final measurement.

- *Experiment 1* – delete activity from sequence relation.
  Event Log –

$$Positive :< a >^{50}$$
$$Negative :< a, b >^{50}$$

  Result Analysis – Figure 6.2 shows the measurements results separately due to the weight change. In Figure 6.2a, with the weight for the existing model increases, precision, accuracy and specificity decrease while the recall keeps 1.0. After checking confusion matrix, we see TP and FN keep the same, FP goes up while TN goes down. The reason is that the higher weight for the existing model causes the higher possibility to keep the existing model during the repair. Due to the deviation of the existing model and actual event log, precision and accuracy goes down. With the weight for positive goes up, all measurements are stable in value 1.0. The reason for this is the TF and TN keeps the same, other values are all 0.0.

-

Test On Parallel

This part shows how the parallel relation of activities is affected by the weights for the existing model, positive and negative instances.

Test On Loop

This part investigated our repair method on activities with loop relation.

Test On Exclusive Choice

This part displays the changes of exclusive choices relation in the model under the different control weights.

6.3.2 Comparison To Other Techniques

This section represents some situations where current repair techniques can't handle properly, while our algorithm gives out an improved repaired model.

Situation 1, unfit part!! added subprocess are too much!! Where the addition of subprocesses and loops are allowed, while the structure changes are impossible, Fahrland's method applies the extension strategy to repair model by adding subprocesses and loops in the procedure. It introduces unseen behavior into the model. However, if the behaviors which are already in the model is unlikely to be removed from the model. One simple example is shown in the following part.

Dee's method is based on Fahrland's method. Deviations are calculated at first and used to build subprocesses for model repair. However, before building subprocesses, it classifies the deviations into positive and negative ones with consideration of trace performance. Only positive deviations are applied to repair model. Different to Fahrland's method, it improves the repaired model performance by limiting the introduced subprocesses. Still, it can't get rid of the defect mentioned before.

Situation 2, For fitted data in the model, can not recognize them!! where overlapped data noise can not be recognized, trace variant with more negative effect is treated as positive and kept in the model, which we should delete them.

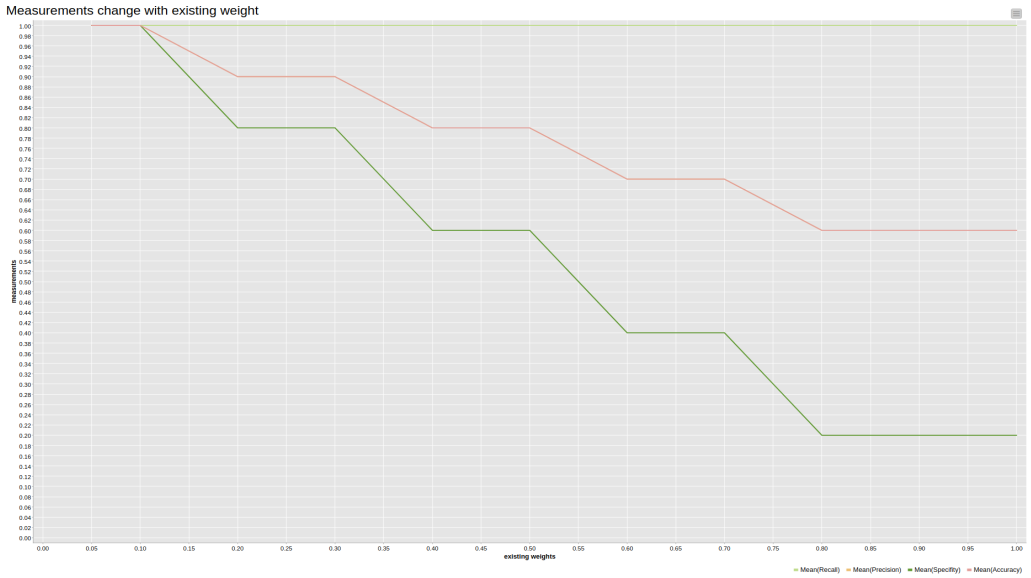
Situation 3, with long-term dependency!! fitted part or new added part!! none of the current techniques can handle this problem yet. Simple examples listed, but will this repeat the last section??

For one exclusive choices, but with long-term dependency detected and added in the model, precision and accuracy increase, since model with long-term dependency blocks the negative information by adding transitions and places to limit activity selection.

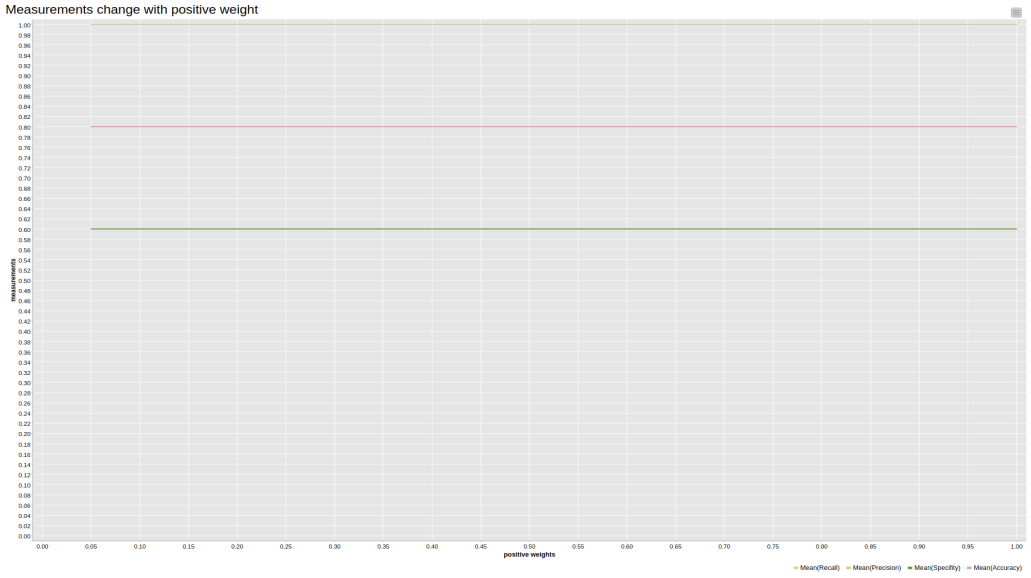
6.3.3 Test On Synthetic Data

Those synthetic data is generated randomly by using the simulation tool plg

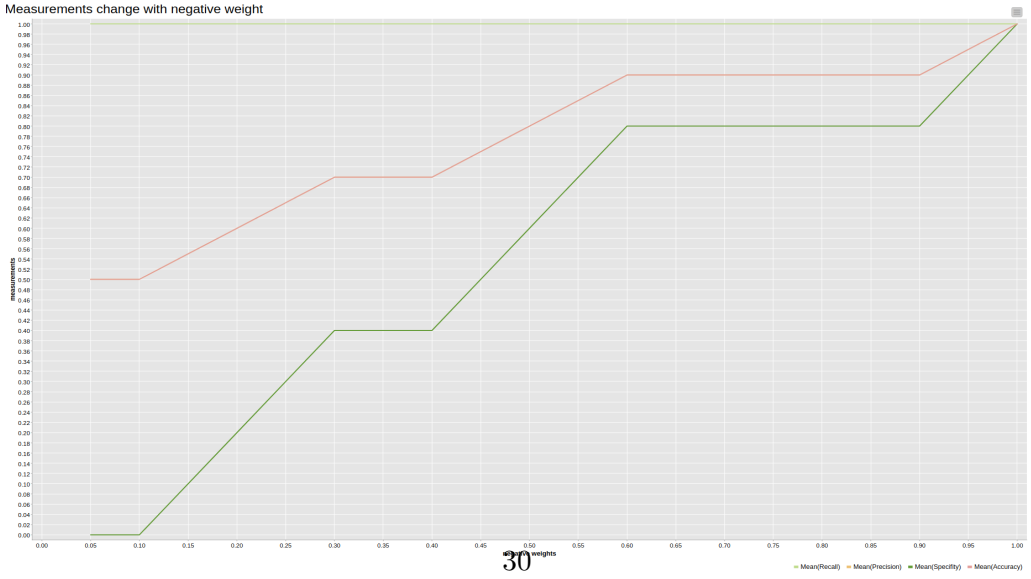
6.3.4 Test On Real life Data



(a) Measurements with the weight for the existing model



(b) Measurements with the weight for positive instances



(c) Measurements with the weight for negative instances

Figure 6.2: example for model change under model repair

Chapter 7

Conclusion

Bibliography

[1] Robin Bergenthum, Jörg Desel, Robert Lorenz, and Sebastian Mauser. Process mining based on regions of languages. In *International Conference on Business Process Management*, pages 375–383. Springer, 2007.

[2] Joos C. A. M. Buijs, Boudewijn F. van Dongen, and Wil M. P. van der Aalst. On the role of fitness, precision, generalization and simplicity in process discovery. In *OTM Conferences*, 2012.

[3] Josep Carmona and Jordi Cortadella. Process discovery algorithms using numerical abstract domains. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):3064–3076, 2014.

[4] Jordi Cortadella, Michael Kishinevsky, Luciano Lavagno, and Alex Yakovlev. Synthesizing petri nets from state-based models. In *Proceedings of IEEE International Conference on Computer Aided Design (ICCAD)*, pages 164–171. IEEE, 1995.

[5] Ana Karla A de Medeiros, Anton JMM Weijters, and Wil MP van der Aalst. Genetic process mining: an experimental evaluation. *Data Mining and Knowledge Discovery*, 14(2):245–304, 2007.

[6] Marcus Dees, Massimiliano de Leoni, and Felix Mannhardt. Enhancing process models to improve business performance: a methodology and case studies. In *OTM Confederated International Conferences" On the Move to Meaningful Internet Systems"*, pages 232–251. Springer, 2017.

[7] Kefang Ding. Incorporatenegativeinformation.

[8] Dirk Fahland and Wil MP van der Aalst. Repairing process models to reflect reality. In *International Conference on Business Process Management*, pages 229–245. Springer, 2012.

[9] Dirk Fahland and Wil MP van der Aalst. Model repair—aligning process models to reality. *Information Systems*, 47:220–243, 2015.

[10] Mahdi Ghasemi and Daniel Amyot. Process mining in healthcare: a systematised literature review. 2016.

[11] Stijn Goedertier, David Martens, Jan Vanthienen, and Bart Baesens. Robust process discovery with artificial negative events. *Journal of Machine Learning Research*, 10(Jun):1305–1340, 2009.

[12] Sander JJ Leemans, Dirk Fahland, and Wil MP van der Aalst. Discovering block-structured process models from event logs-a constructive approach. In *International conference on applications and theory of Petri nets and concurrency*, pages 311–329. Springer, 2013.

[13] Hernan Ponce-de León, Josep Carmona, and Seppe KLM vanden Broucke. Incorporating negative information in process discovery. In *International Conference on Business Process Management*, pages 126–143. Springer, 2016.

[14] Eindhoven Technical University. © 2010. Process Mining Group. Prom introduction.

[15] Wil Van Der Aalst. *Process mining: discovery, conformance and enhancement of business processes*, volume 2. Springer, 2011.

[16] Wil Van der Aalst. Data science in action. In *Process Mining*, pages 3–23. Springer, 2016.

[17] Wil van der Aalst. *Process Mining: Data Science in Action*. Springer Publishing Company, Incorporated, 2nd edition, 2016.

[18] Wil Van der Aalst, Ton Weijters, and Laura Maruster. Workflow mining: Discovering process models from event logs. *IEEE Transactions on Knowledge and Data Engineering*, 16(9):1128–1142, 2004.

[19] Jan Martijn EM Van der Werf, Boudewijn F van Dongen, Cor AJ Hurkens, and Alexander Serebrenik. Process discovery using integer linear programming. In *International conference on applications and theory of petri nets*, pages 368–387. Springer, 2008.

[20] Boudewijn F Van Dongen, AK Alves De Medeiros, and Lijie Wen. Process mining: Overview and outlook of petri net discovery algorithms. In *Transactions on Petri Nets and Other Models of Concurrency II*, pages 225–242. Springer, 2009.

[21] Seppe KLM vanden Broucke, Jochen De Weerdt, Jan Vanthienen, and Bart Baesens. Determining process model precision and generalization with weighted artificial negative events. *IEEE Transactions on Knowledge and Data Engineering*, 26(8):1877–1889, 2014.

[22] Anton JMM Weijters and Wil MP Van der Aalst. Rediscovering workflow models from event-based data using little thumb. *Integrated Computer-Aided Engineering*, 10(2):151–162, 2003.