INDUCTIVE CONTEXT-AWARE PROCESS DISCOVERY

Roee Shraga, Avigdor Gal, Dafna Schumacher – Technion
Arik Senderovich - University of Toronto
Matthias Weidlich - Humboldt-Universität zu Berlin
## Motivation

* Each case in ordered by time of day.

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Resource</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>W</td>
<td>11:10</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>X</td>
<td>18:00</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>X</td>
<td>18:30</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>X</td>
<td>10:00</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>X</td>
<td>10:00</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>W</td>
<td>10:10</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
<td>X</td>
<td>09:00</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
<td>X</td>
<td>09:00</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
<td>W</td>
<td>09:05</td>
</tr>
</tbody>
</table>

Diagram:

- **A** ➔ **D**
  - $time \geq 14:05$
- **B** ➔ **C**
  - $time < 14:05$

<table>
<thead>
<tr>
<th>Activity</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Scribble notes</td>
</tr>
<tr>
<td>B</td>
<td>Send notes to assistant</td>
</tr>
<tr>
<td>C</td>
<td>Type notes</td>
</tr>
<tr>
<td>D</td>
<td>Voice-to-text typing</td>
</tr>
</tbody>
</table>
Outline

- Preliminaries
  - Inductive Miner
  - K-means clustering
  - Silhouette
- Inductive Context-Aware Process Discovery
  - CaT - Context-Aware Tree
  - CaDi - Context-Aware Discovery
- Evaluation
- Conclusions
Event-logs

- $e$ is an event
  - Each event $e$ has an Activity, $e.a$
  - and a set of attributes $e.x_1, e.x_2, \ldots, e.x_p$ (context in denoted by the set of attributes $\chi$).
- $\sigma = \langle e_1, \ldots, e_n \rangle$ is a trace
- $\pi_\oplus (L)$ is a log partition

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Resource</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>W</td>
<td>11:10</td>
</tr>
</tbody>
</table>

$x_1, x_2$
**Process Trees**

- A rooted tree that represents a process.
- Each leaf node is an activity and all non-leaf nodes are control-flow operators.
- The process tree is given as a set of activity sequence: \( \{\langle A, B, C \rangle, \langle A, D \rangle\} \)

<table>
<thead>
<tr>
<th>Operator</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>→</td>
<td>Sequence</td>
</tr>
<tr>
<td>( \times )</td>
<td>Exclusive choice</td>
</tr>
<tr>
<td>( \land )</td>
<td>Concurrency</td>
</tr>
<tr>
<td>( \triangledown )</td>
<td>Loop</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>B</td>
</tr>
<tr>
<td>4</td>
<td>C</td>
</tr>
</tbody>
</table>
The IM algorithm (Leemans et al. in [1]) recursively applies a select function to construct a process tree as follows:

Select operator $\oplus$ and log partitioning $\pi_{\oplus}(L)$

Recurse on each $l \in \pi_{\oplus}(L)$

We use the Inductive Miner to capture control-flow.

---

K-means Clustering

- Each event $e$ can be seen as a $p$-dimensional feature vector $(e.x_1, \ldots, e.x_p)$.
- We use K-means Clustering to capture context.

<table>
<thead>
<tr>
<th>Case</th>
<th>Activity</th>
<th>Resource</th>
<th>Time</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>B</td>
<td>X</td>
<td>11:00</td>
</tr>
<tr>
<td>1</td>
<td>C</td>
<td>W</td>
<td>11:10</td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td>X</td>
<td>18:00</td>
</tr>
<tr>
<td>2</td>
<td>D</td>
<td>X</td>
<td>18:30</td>
</tr>
<tr>
<td>3</td>
<td>A</td>
<td>X</td>
<td>10:00</td>
</tr>
<tr>
<td>3</td>
<td>B</td>
<td>X</td>
<td>10:00</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>W</td>
<td>10:10</td>
</tr>
</tbody>
</table>
**K-means – Mixed Distance Measure**

- $\chi_N \subseteq \chi$ numeric attributes. The distance computed by their **Euclidean** distance.
- $\chi_C \subseteq \chi$ categorical attributes. The distance computed by:

$$d_e(e.x, e'.x) = \begin{cases} 0, & \text{if } e.x = e'x, \\ 1, & \text{if } e.x \neq e'x \end{cases}$$

- The mixed measure:

$$d(e, e') = \omega \sqrt{\sum_{x \in \chi_N} (e.x - e'.x)^2 + (1 - \omega) \sum_{x \in \chi_C} d_e(e.x, e'.x)}$$

$$\omega = \frac{|\chi_N|}{|\chi|}$$
K-means Clustering
Silhouette

Silhouette ($\psi_s$) is quality measure of a split of the events of a log.

It given by quantification the separation between clusters.

$-1 \leq \psi_s(e) \leq 1$

The silhouette coefficient $\psi_s(e)$ is defined by:

$$
in(e) = \frac{\sum_{e' \in E_j \setminus \{e\}} d(e, e')}{|E_j|}$$

$$
out(e) = \min_{1 \leq i \leq K, i \neq j} \frac{\sum_{e \in E_i} d(e, e')}{|E_i|}$$

$$
\psi_s(e) = \frac{out(e) - in(e)}{\max(in(e), out(e))}
$$

$$
\psi_s(\{E_1, ..., E_K\}) = \frac{1}{K} \sum_{1 \leq j \leq K} \frac{\sum_{e \in E_j} \psi_s(e)}{|E_j|}
$$
Outline

✓ Preliminaries
  ✓ Inductive Miner
  ✓ K-means clustering
  ✓ Silhouette

➢ Inductive Context-Aware Process Discovery
  ▪ CaT- Context-Aware Tree
  ▪ CaDi - Context-Aware Discovery

➢ Evaluation

➢ Conclusions
CaT- Context Aware Process Trees

- CaTs extend process trees with an operator that carries data semantics.
- Constraint operator (*)

![Diagram of CaT process trees](image-url)
CaT- Context Aware Process Trees

- We set $K = 2$, which allows us to use the boundary between the clusters to annotate the Constraint operator.
**CaDi- Context Aware Discovery Overview**

- **Inductive Miner**
- **Control-flow**
- **Context**
- **K-means**

\[
\psi_S \rightarrow \text{CaDi} \rightarrow \text{CaT}
\]
Algorithm 1 Context-aware Process Discovery (CaDi)

1: procedure CaDi(event log $L$, quality threshold $\delta$)
2:   Create a CaT $T$
3:   \[ \text{for } \text{each event } o \text{ in } L \text{ do} \]
4:     \[ \text{if } \text{context aware \textbf{THEN}} \]
5:       \[ \text{context} \]
6:     \[ \text{end if} \]
7:   \[ \text{end for} \]
8:   return $T$
9: end procedure
CaDi- Context Aware Discovery

Input: $L, \delta$

Clustering: 2-means

$\psi_s$

$\psi_s > \delta$ \hspace{1cm} $\psi_s \leq \delta$

Add $\ast$

Add IM operator

Split $L$ to $L'$

For each $L'$

if: $\forall e, e' \in E_{L'} : e.a = e'.a$

Add a
Outline

✓ Preliminaries
  ✓ Inductive Miner
  ✓ K-means clustering
  ✓ Silhouette

✓ Inductive Context-Aware Process Discovery
  ✓ CaT - Context-Aware Tree
  ✓ CaDi - Context-Aware Discovery

➢ Evaluation
➢ Conclusions
Evaluation – Experimental Setup

➢ Quality:

- **Multi-perspective**: Fitness ($\psi_{Fit}(L,T)$) and precision ($\psi_{Pre}(L,T)$) [3]
- Generalization ($\psi_{Gen}(L,T)$) in a training and validation fashion.
- $\psi_{Tot}(L,T) = \frac{1}{3} \psi_{Fit}(L,T) + \frac{1}{3} \psi_{Pre}(L,T) + \frac{1}{3} \psi_{Gen}(L,T)$

➢ Implementation:

- **Python code**
  - NLTK for K-means
  - EDU-ProM for IM [3]
- Produce a Data Petri net (DPN)
- Evaluate using multi-perspective process explorer in ProM

➢ Baseline:

- Inductive Mining followed by Decision Mining (IM&DM) [4].

---

Evaluation – Datasets

- Three synthetic event logs:
  - Randomized process tree [5].
  - 500 traces, with an average length of 3 events.
  - Three randomly sampled attributes, namely: time-of-day, resource, and duration.

- Real-world hospital log:
  - A single day of event data from an outpatient cancer hospital in the United States.
  - 500 treatment traces that consist of a total of 1,200 events.
  - One numeric feature and two categorical features.

Evaluation – Results

- The multi-perspective quality measures as function of $\delta$ over synthetic log.

$\psi_s > \delta \quad \Rightarrow \quad \psi_s \leq \delta$

Add $\ast$  
Add IM operator
Evaluation – Results

- The overall performance in terms of Fitness ($\psi_{Fit}$), Precision ($\psi_{Pre}$), Generalization ($\psi_{Gen}$) and Total Quality ($\psi_{Tot}$) obtained by CaT.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Method</th>
<th>$\psi_{Fit}$</th>
<th>$\psi_{Pre}$</th>
<th>$\psi_{Gen}$</th>
<th>$\psi_{Tot}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Log1</td>
<td>CaDi (0.4)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>IM&amp;DM</td>
<td>0.993</td>
<td>1.0</td>
<td>0.977</td>
<td>0.99</td>
</tr>
<tr>
<td>Log2</td>
<td>CaDi (0.75)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>IM&amp;DM</td>
<td>0.875</td>
<td>1.0</td>
<td>0.869</td>
<td>0.915</td>
</tr>
<tr>
<td>Log3</td>
<td>CaDi (0.85)</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>IM&amp;DM</td>
<td>0.75</td>
<td>1.0</td>
<td>0.75</td>
<td>0.833</td>
</tr>
<tr>
<td>Hospital</td>
<td>CaDi (0.45)</td>
<td>0.917</td>
<td>0.957</td>
<td>0.868</td>
<td>0.914</td>
</tr>
<tr>
<td></td>
<td>IM&amp;DM</td>
<td>0.826</td>
<td>0.785</td>
<td>0.838</td>
<td>0.816</td>
</tr>
</tbody>
</table>
Evaluation – Results

- Comparison of CaDi (blue) and IM&DM (red) in terms of data (x-axis) and control-flow (y-axis) deviations over the Hospital dataset. Larger markers indicate higher frequency of the respective values.
Outline

✓ Preliminaries
  ✓ Inductive Miner
  ✓ K-means clustering
  ✓ Silhouette
✓ Inductive Context-Aware Process Discovery
  ✓ CaT - Context-Aware Tree
  ✓ CaDi - Context-Aware Discovery
✓ Evaluation
  ➢ Conclusions
Conclusions

- A novel discovery framework that considers both control-flow and context, simultaneously.
- A discovery algorithm, CaDi, which combines Inductive Mining and K-means clustering to generate a context aware process tree, CaT.
- We validate the superiority of CaDi, using multi-perspective alignments.
Thank You For Listening!

Any Questions?

Personal URL:
https://sites.google.com/view/roee-shraga

Contact me at:
shraga89@campus.technion.ac.il

Code:
https://github.com/dafna-s/Inductive-Context-aware-Process-Discovery