Discovering Predictive Dependencies on Multi-Temporal Relations

Beatrice Amico, Carlo Combi, Pietro Sala, Romeo Rizzi

Department of Computer Science, University of Verona

Athens, TIME 2023
Table of Contents

1 Introduction

2 A 3-window model for the interpretation of predictive temporal data
   - A 3-window model
   - Multi-temporal relational model
   - Time-frame tuple consistency
   - Predictive functional dependencies
   - Discovering Approximate PFD

3 The computational aspects of APFDs

4 Experimental evaluation
   - Application domain
   - Dataset and preprocessing
   - Results

5 Conclusions
Introduction

2 A 3-window model for the interpretation of predictive temporal data
   ■ A 3-window model
   ■ Multi-temporal relational model
   ■ Time-frame tuple consistency
   ■ Predictive functional dependencies
   ■ Discovering Approximate PFD

3 The computational aspects of APFDs

4 Experimental evaluation
   ■ Application domain
   ■ Dataset and preprocessing
   ■ Results

5 Conclusions
**Motivation**: temporal patterns represent an explainable way to study the intrinsic data dependencies. Mining functional dependencies can be fruitfully exploited to improve prediction, often related to ML models.

**Goal**: we propose a temporally-oriented data mining framework to support the prediction based on the identification of recurring temporal patterns, the Approximate Temporal Predictive Functional Dependencies (APFDs), within a 3-window-based temporal framework.
An FD is composed of the antecedent ($X$) and the consequent ($Y$). Informally, for all the couples of tuples $t$ and $t'$ showing the same value(s) on $X$, the corresponding value(s) on $Y$ are identical.

$$X \rightarrow Y$$

Through the use of functional dependencies, we can express concepts such as: “for each drug with a given symptom the disease does not change”:

$$Drug, Symptom \rightarrow Disease.$$
When we add temporal extensions to the atemporal functional dependencies, we talk about temporal functional dependency (TFD).

Through the use of temporal functional dependencies, we can express concepts such as “for each drug with a given symptom the received diagnosis does not change, over a time windows of 10 days”:

\[ [10 \text{ days}] \ Drug, \ Symptom \Rightarrow Diagnosis \]
Approximate Functional Dependencies

An AFD $f$ requires the FD to be satisfied by most tuples of relation $w$. It allows a very small portion of tuples of $w$ to violate the dependency.

If this portion is less than or equal to the satisfaction threshold $\varepsilon$, $f$ is approximately satisfied on $s$.

Prediction?
Table of Contents

1 Introduction

2 A 3-window model for the interpretation of predictive temporal data
   - A 3-window model
   - Multi-temporal relational model
   - Time-frame tuple consistency
   - Predictive functional dependencies
   - Discovering Approximate PFD

3 The computational aspects of APFDs

4 Experimental evaluation
   - Application domain
   - Dataset and preprocessing
   - Results

5 Conclusions
A 3-window model for the interpretation of predictive temporal data

We generalize an approach based on three (possibly moving) time windows:

- **Observation window**: a time interval where the information is collected;
- **Waiting window**: the minimum time interval required to act in order to prevent the event in the prediction window;
- **Prediction window**: the time interval when the predicted event occurs.
**Anchored and unanchored windows**

**Anchored time windows** represent specific periods of the considered time axis.

**Unanchored time windows** represent windows that ”move” through the time axis, constraining only the distance between the considered data.
A second distinction for the time windows, which may provide different results for prediction is:

- **fixed-length**: OW, WW, PW have a fixed length without any further constraint related to the temporal position of data inside them;

- **variable-length**: OW, WW, PW end with the last time point associated with the data to consider in the window.
A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

<table>
<thead>
<tr>
<th>#</th>
<th>Patient</th>
<th>HR</th>
<th>VT</th>
<th>SpO₂</th>
<th>VT</th>
<th>Drug</th>
<th>VT</th>
<th>AKI</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daisy</td>
<td>High</td>
<td>19</td>
<td>High</td>
<td>21</td>
<td>Aspirin</td>
<td>23</td>
<td>False</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>High</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>Medium</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Daisy</td>
<td>Medium</td>
<td>5</td>
<td>Medium</td>
<td>7</td>
<td>Indapamid</td>
<td>9</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Stevie</td>
<td>Medium</td>
<td>4</td>
<td>Medium</td>
<td>7</td>
<td>Metolazone</td>
<td>8</td>
<td>True</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Indapamid</td>
<td>7</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
A multi-temporal relation \( mrt \) is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

<table>
<thead>
<tr>
<th>#</th>
<th>Patient</th>
<th>HR(^1)</th>
<th>VT(^1)</th>
<th>SpO(_2)(^2)</th>
<th>VT(^2)</th>
<th>Drug(^3)</th>
<th>VT(^3)</th>
<th>AKI</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daisy</td>
<td>High</td>
<td>19</td>
<td>High</td>
<td>21</td>
<td>Aspirin</td>
<td>23</td>
<td>False</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>High</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>Medium</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Daisy</td>
<td>Medium</td>
<td>5</td>
<td>Medium</td>
<td>7</td>
<td>Indapamid</td>
<td>9</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Stevie</td>
<td>Medium</td>
<td>4</td>
<td>Medium</td>
<td>7</td>
<td>Metolazone</td>
<td>8</td>
<td>True</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Indapamid</td>
<td>7</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
A multi-temporal relation *mrt* is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

<table>
<thead>
<tr>
<th>#</th>
<th>Patient</th>
<th>HR&lt;sup&gt;1&lt;/sup&gt;</th>
<th>VT&lt;sup&gt;1&lt;/sup&gt;</th>
<th>SpO&lt;sub&gt;2&lt;/sub&gt;&lt;sup&gt;2&lt;/sup&gt;</th>
<th>VT&lt;sup&gt;2&lt;/sup&gt;</th>
<th>Drug&lt;sup&gt;3&lt;/sup&gt;</th>
<th>VT&lt;sup&gt;3&lt;/sup&gt;</th>
<th>AKI</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daisy</td>
<td>High</td>
<td>19</td>
<td>High</td>
<td>21</td>
<td>Aspirin</td>
<td>23</td>
<td>False</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>High</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>Medium</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Daisy</td>
<td>Medium</td>
<td>5</td>
<td>Medium</td>
<td>7</td>
<td>Indapamid</td>
<td>9</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Stevie</td>
<td>Medium</td>
<td>4</td>
<td>Medium</td>
<td>7</td>
<td>Metolazone</td>
<td>8</td>
<td>True</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Indapamid</td>
<td>7</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>36</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
A multi-temporal relation $mrt$ is characterized by multiple valid times. Each tuple of such relation represents a piece of history of a given entity, through the values of attributes holding at different (valid) times.

<table>
<thead>
<tr>
<th>#</th>
<th>Patient</th>
<th>HR$^1$</th>
<th>VT$^1$</th>
<th>SpO$_2$$^2$</th>
<th>VT$^2$</th>
<th>Drug$^3$</th>
<th>VT$^3$</th>
<th>AKI</th>
<th>VT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Daisy</td>
<td>High</td>
<td>19</td>
<td>High</td>
<td>21</td>
<td>Aspirin</td>
<td>23</td>
<td>False</td>
<td>28</td>
</tr>
<tr>
<td>2</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>High</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>3</td>
<td>Daisy</td>
<td>Low</td>
<td>2</td>
<td>Medium</td>
<td>4</td>
<td>Aspirin</td>
<td>6</td>
<td>False</td>
<td>12</td>
</tr>
<tr>
<td>4</td>
<td>Daisy</td>
<td>Medium</td>
<td>5</td>
<td>Medium</td>
<td>7</td>
<td>Indapamid</td>
<td>9</td>
<td>False</td>
<td>18</td>
</tr>
<tr>
<td>5</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>17</td>
</tr>
<tr>
<td>6</td>
<td>Luke</td>
<td>Low</td>
<td>7</td>
<td>High</td>
<td>8</td>
<td>Ibuprofen</td>
<td>12</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>18</td>
</tr>
<tr>
<td>8</td>
<td>Luke</td>
<td>Medium</td>
<td>9</td>
<td>High</td>
<td>13</td>
<td>Sulindac</td>
<td>14</td>
<td>True</td>
<td>21</td>
</tr>
<tr>
<td>9</td>
<td>Stevie</td>
<td>Medium</td>
<td>4</td>
<td>Medium</td>
<td>7</td>
<td>Metolazone</td>
<td>8</td>
<td>True</td>
<td>13</td>
</tr>
<tr>
<td>10</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Indapamid</td>
<td>7</td>
<td>False</td>
<td>8</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>36</td>
<td>Stevie</td>
<td>High</td>
<td>1</td>
<td>Low</td>
<td>2</td>
<td>Aspirin</td>
<td>5</td>
<td>False</td>
<td>25</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
Given a multi-temporal relation $mtr$, now we are interested in verifying which tuples are “fine” with, or “contained” in, a given time frame.

We are interested in eliciting those tuples having the $k$ observation-related valid times contained in the observation window $OW$, and the last valid time in the prediction window $PW$.

We will call them **consistent** with the considered time frame.
Time-frame tuple consistency

Given a tuple \( t \) of a multi-temporal relation \( mrt \), we say that \( t \) is time-frame consistent if the formula \( \Theta(t, \alpha, m, [i_1, i_2]) \) holds.

There exist different possible formulas according to the different choice of variables:

- \( \alpha \): anchored or unanchored time frame;
- \( m \): fixed or flex modality;
- \([i_1, i_2]\): VT attribute range within the observation window.
Time-frame tuple consistency

Given a tuple $t$ of a multi-temporal relation $mrt$, we say that $t$ is time-frame consistent if the formula $\Theta(t, \alpha, m, [i_1, i_2])$ holds.

There exist different possible formulas according to the different choice of variables:

- $\alpha$: anchored or unanchored time frame;
- $m$: fixed or flex modality;
- $[i_1, i_2]$: VT attribute range within the observation window.

$$
\Theta(t, \alpha, 'flex', [i_1, i_2]) \equiv t[\overrightarrow{VT}^i_2] - t[\overrightarrow{VT}^i_1] \leq OW \land t[\dot{VT}] - t[\overrightarrow{VT}^i_2] > WW \land t[\dot{VT}] - t[\overrightarrow{VT}^i_2] < WW + PW
$$
Discovering Predictive Dependencies on Multi-Temporal Relations

**General idea:** propose a general framework allowing the definition of “specialized” functional dependencies having:

- the **antecedent** composed of a set of attributes related to “past” properties, called predictive attributes, ordered according to the corresponding valid times;

- the **consequent** composed of a set of attributes related to “future” properties, called predicted attributes.
Predictive Functional Dependency (PFD)

**Definition**

Given:

- an mt-relation schema $MTR(Z \bar{U}^1 \bar{U}^2 .. \bar{U}^k \hat{U} \cup \{\bar{V}T^1, \bar{V}T^2, .., \bar{V}T^k, \bar{V}T\})$ where $\bar{U}_i$ is a set of attributes representing properties of an entity and $Z$ are the identification attributes;
- a time frame;
- a modality $m \in \{'flex', 'fixed'\}$.

A **Predictive Functional Dependency** is expressed as:

$$SP^hQ^i...R^j \xrightarrow{\alpha, m} \hat{Y} \quad \text{with} \quad 1 \leq h < i < ... < j \leq k$$

where $S \subseteq Z, P^h \subseteq \bar{U}^h, Q^i \subseteq \bar{U}^i, R^j \subseteq \bar{U}^j$ and $\hat{Y} \subseteq \hat{U}$ is the predicted attribute set.
Discovering Approximate PFD (APFD)

We need to deal with some kind of approximation, as it could happen that some PFDs hold on a subset of tuples of the time-frame relation view, we consider.

In other words, we require a PFD $f$ to be satisfied by most tuples of the TF-view $w$, $w \subseteq mtr$.

A very small portion of tuples of $w$ is allowed to violate the dependency. In the context of APFDs, we consider three error measures: $G_3$, $H_3$, $J_3$. 
Approximation: Error $G_3$

Given a TF-view $w = TFv(mtr, \alpha, m, [1, k])$ of an mt-relation $mtr$, and a PFD $S\overline{P^hQ^i...R^j} \xrightarrow{\alpha,m} \dot{Y}$, where $S \subseteq Z$, $\overline{P^h} \subseteq \overline{U^h}$, $\overline{Q^i} \subseteq \overline{U^i}$, $\overline{R^j} \subseteq \overline{U^j}$ and $\dot{Y} \subseteq \dot{U}$, and any relation $s \subseteq w$, such that $s \models E_{\alpha,m} S\overline{P^hQ^i...R^j} \rightarrow \dot{Y}$, we define three errors:

- $G_3$ considers the minimum number of tuples in $w$ to be deleted to obtain a relation $s$ where the given FD holds.

$G_3$ is expressed as:

$$G_3 = |w| - |s|$$

The related scaled measurement $g_3$ is defined as:

$$g_3 = \frac{G_3}{|w|}$$
Approximation: Error $H_3$

- $H_3$ is focused on the number of entities that we accept to discard for the sake of the PFD (for example disregard data of entities with a very low number of tuples, which could create noise in our dataset).

$H_3$ is expressed as:

$$H_3 = |\{t[Z] \mid \exists t \in w\}| - |\{t[Z] \mid \exists t \in s\}|$$

The related \textit{scaled measurement} $h_3$ is defined as:

$$h_3 = \frac{H_3}{|\{t[Z] \mid \exists t \in w\}|}$$
Approximation:: Error $J_3$

- $J_3$ considers the number of tuples for each entity we accept to discard to satisfy the PFD. It ensures to maintain enough “consistent” information for each entity.

$J_3$ is expressed as:

$$J_3 = \max_{(v \in \{ t[Z] | t \in s \})} \{ |w_v| - |s_v| \}$$

$w_v \equiv \{ t[Z] | t \in w \land t[Z] = v \}$ and $s_v \equiv \{ t[Z] | t \in s \land t[Z] = v \}$

The related scaled measurement $j_3$ is defined as follows:

$$j_3 = \max_{(v \in \{ t[Z] | t \in s \})} \left\{ \frac{|w_v| - |s_v|}{|w_v|} \right\}$$
Definition (Approximate Predictive Functional Dependency (APFD))

Given a TF-view $w = TFv(mtr, \alpha, m, [1, k])$ of an mt-relation $mtr$ with schema $Z \cup \{ V^1, V^2, \ldots, V^k, \tilde{V} \} \cup \{ U^1\cup \{ V^1, V^2, \ldots, V^k, \tilde{V} \} \},$ $w$ fulfills the APFD

$$S \bar{P}^h Q^i \ldots \bar{R}^j \xrightarrow[\alpha, m]{} \dot{Y}$$

(written as $w \models^{E, \alpha, m} S \bar{P}^h Q^i \ldots \bar{R}^j \xrightarrow{} \dot{Y}$), where $\varepsilon = \langle \varepsilon_g, \varepsilon_h, \varepsilon_j \rangle$ and $S \subseteq Z, \bar{P}^h \subseteq \hat{U}^h, Q^i \subseteq \hat{U}^i, \bar{R}^j \subseteq \hat{U}^j, \dot{Y} \subseteq \hat{U},$ if a relation $s \subseteq w$ exists such that $s \models^{E, \alpha, m} S \bar{P}^h Q^i \ldots \bar{R}^j \rightarrow \dot{Y}$ with $g_3 \leq \varepsilon_g \land h_3 \leq \varepsilon_h \land j_3 \leq \varepsilon_h.$

$\varepsilon_g, \varepsilon_h, \varepsilon_j$ are the maximum acceptable errors defined by the user for $g_3, h_3, j_3$ respectively.
1 Introduction

2 A 3-window model for the interpretation of predictive temporal data
   - A 3-window model
   - Multi-temporal relational model
   - Time-frame tuple consistency
   - Predictive functional dependencies
   - Discovering Approximate PFD

3 The computational aspects of APFDs

4 Experimental evaluation
   - Application domain
   - Dataset and preprocessing
   - Results

5 Conclusions
To discuss the complexity of checking an APFD, it is enough to consider a relation having:

- a single attribute \( A \) representing the antecedent;
- the predicted attribute \( \hat{B} \);
- a single attribute \( Z \) representing the entity attribute.

The domain of all attributes is \( \mathbb{N} \) or a subset of it (the predicted values for \( \hat{B} \) will be either 0 or 1, to represent boolean values).

Thus, we will consider a relation \( w \) with schema

\[ W(A, \hat{B}, Z) \]
The (data) complexity of deriving an APFD

Given a relation $w \subset \mathbb{N}^3$, a natural number $0 \leq k < |w|$, and a natural number $0 \leq h < |\pi_Z(w)|$ determine whether or not $w$ admits a conflict resolution of order $(k,h)$.

$k$ represents the threshold $G_3$.

$h$ represents the threshold $H_3$.

We prove that the problem of verifying any APFD even only considering $H_3$ is NP-Hard. (Proof by reduction from an already known problem \(^1\).)

We reduced the problem in hand to a general $3SAT$ problem, showing that checking an APFD considering all the three thresholds belongs to the class $NP$.

An instance of $3SAT$ problem is a logical formula formed by a conjunction of disjunctive clauses, where each clause has exactly 3 literals.

$$\left( X_1 \lor X_2 \lor X_3 \right) \land \left( X_4 \lor X_5 \lor X_6 \right) \land \left( X_7 \lor X_8 \lor X_9 \right)$$
After proving that verifying any APFD even only considering $H_3$ is NP-Hard, we propose a deterministic algorithm that could stop the analysis of a relation, as soon as it verifies that the relation cannot satisfy the given APFD.

**General idea**: searching for a solution considering one tuple at a time, until it is possible to generate a solution, which satisfies the selected thresholds.
Table of Contents

1 Introduction

2 A 3-window model for the interpretation of predictive temporal data
   ■ A 3-window model
   ■ Multi-temporal relational model
   ■ Time-frame tuple consistency
   ■ Predictive functional dependencies
   ■ Discovering Approximate PFD

3 The computational aspects of APFDs

4 Experimental evaluation
   ■ Application domain
   ■ Dataset and preprocessing
   ■ Results

5 Conclusions
Intensive care unit

- Physicians have the access to a large quantity of data for each patient, derived from the continuous monitoring.
- **Timing** is a fundamental part: anticipation of the illness onset, worsening of clinical condition or the diagnosis moment.
- It could be difficult to identify knowledge for clinical decisions: data mining techniques are useful to identify the most significant information.
AKI is a syndrome characterized by sudden kidney failure (high values of creatinine and low urine output) with a rapid progression.
Experimental evaluation: dataset and preprocessing

MIMIC III

50.711 subjects

Tables: D_ITEMS, D_LABITEMS, PATIENTS, ICUSTAYS, PRESCRIPTIONS, LABEVENTS, CHARTEVENTS

KDIGO criteria

Five measures: creatinine, administered drug (diuretics, non-steroidal anti-inflammatory drugs (NSAID), radiocontrast agents, and angiotensin), respiratory rate, oxygen saturation, and diastolic blood pressure

Numerical parameters categorization

Two unanchored time frames:
- OW 72, WW 12, PW 36 (hours)
- OW 120, WW 12, PW 36 (hours)
Experimental evaluation: dataset and preprocessing

Three different TF-views

**TF-view #1**
- Four states of the same measure (serum creatinine) to build a sequence where any value is the next of the preceding one (if any),
- OW 72, WW 12, PW 36 (hours)
- 2546 subjects (1878 controls, 668 cases in 3839 rows)

**TF-view #2**
- Four states of the same measure (administered drugs) to build a sequence where any value is the next of the preceding one (if any),
- OW 120, WW 12, PW 36 (hours)
- 148 subjects (109 controls, 39 cases in 1047 rows)

**TF-view #3**
- Four states each one related to a different measure (administered drug, diastolic blood pressure, respiratory rate, oxygen saturation) with $\overline{V}_T^k = \overline{V}_T^{k-1} + 1 \text{ for } k = 1, \ldots, 3$
- OW 120, WW 12, PW 36 (hours)
- 413 subjects (305 controls, 108 cases in 193,173 rows)

Computing APFDs

APFDs
We report some of the APFDs obtained through the algorithm, with the corresponding error thresholds.

<table>
<thead>
<tr>
<th>APFD</th>
<th>$\varepsilon_g$</th>
<th>$\varepsilon_h$</th>
<th>$\varepsilon_j$</th>
<th>TF-view</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Creat^1, Creat^3 \rightarrow AKI$</td>
<td>27.45%</td>
<td>27%</td>
<td>50%</td>
<td>#1</td>
</tr>
<tr>
<td>$Creat^1, Creat^4 \rightarrow AKI$</td>
<td>27.45%</td>
<td>27%</td>
<td>50%</td>
<td>#1</td>
</tr>
<tr>
<td>$Drug^1, Drug^2, Drug^4 \rightarrow AKI$</td>
<td>21%</td>
<td>30%</td>
<td>50%</td>
<td>#2</td>
</tr>
<tr>
<td>$Drug^1, Drug^2, Drug^4 \rightarrow AKI$</td>
<td>21%</td>
<td>30%</td>
<td>80%</td>
<td>#2</td>
</tr>
<tr>
<td>$Drug^1, Drug^2, Drug^3 \rightarrow AKI$</td>
<td>21%</td>
<td>30%</td>
<td>80%</td>
<td>#2</td>
</tr>
<tr>
<td>$Drug^1, Drug^3, Drug^4 \rightarrow AKI$</td>
<td>21%</td>
<td>30%</td>
<td>80%</td>
<td>#2</td>
</tr>
<tr>
<td>$Drug^1, RespRate^3 \rightarrow AKI$</td>
<td>10%</td>
<td>51%</td>
<td>75%</td>
<td>#3</td>
</tr>
<tr>
<td>$RespRate^3 \rightarrow AKI$</td>
<td>30%</td>
<td>75%</td>
<td>75%</td>
<td>#3</td>
</tr>
<tr>
<td>$Drug^1 \rightarrow AKI$</td>
<td>30%</td>
<td>75%</td>
<td>75%</td>
<td>#3</td>
</tr>
<tr>
<td>$Spo_2^4 \rightarrow AKI$</td>
<td>30%</td>
<td>75%</td>
<td>75%</td>
<td>#3</td>
</tr>
</tbody>
</table>
We proposed a methodology for deriving a new kind of approximate temporal functional dependencies, called **Approximate Predictive Functional Dependencies**.

- A formal 3-window model to derive the APFDs;
- The computational aspects of deriving an APFD;
- The application to real clinical data, specifically to MIMIC III dataset.
Thank you for your attention!
Extra slides
The applicability of our framework

We can apply the entire framework in every domain where the prediction task could be an interesting task.

The usefulness of the 3-widow model is tied to two aspects:

• the final goal related to the problem in hand;
• the nature of the predicted event.
Why a 3-window model

The nature of the predicted event.

The waiting window is used to anticipate an action in order to prevent a future event. It is important to underline that not every type of events could be prevented.

Example in medicine: Diabetes diagnosis, we cannot prevent this diagnosis, because it’s a fact that simply happened at a certain point, and we cannot avoid it.

AKI, Sepsis, Covid-19 are diseases that imply a possible deterioration or improvement of the patient status. So in this case, the waiting window could be use to anticipate as soon as possible the diagnosis, preventing the deterioration.
Why a 3-window model

The final goal related to the problem in hand.

Using again the diabetes diagnosis. Suppose to have a database that records EHR from a childhood diabetes center.

A way to use our model could be consider the final goal to study all the different temporal events such as specialist visit, hospitalization in the emergency department, in order to anticipate the start of the cure of these patients.

In this case it is not possible to prevent an event (diabetes) that is unavoidable, but we can the anticipate the moment of the diagnosis, the start of the treatment in order to alleviate the long term side-effects.
A PFD holds on an mt-relation \( mtr \) with schema \( MTR \) in a timeframe \( TF \) with modality \( m \), with a restricted or extended range semantics (denoted as \( mtr \models^R_{\alpha,m} \) or \( mtr \models^E_{\alpha,m} \)) iff:

\[
\forall t, t' \in mtr((t[S^hP^i ..R^j] = t'[S^hP^i ..R^j] \land \Theta(t, \alpha, m, [h,j]) \land \Theta(t', \alpha, m, [h,j])) \rightarrow t[\dot{Y}] = t'[\dot{Y}])
\]

or

\[
\forall t, t' \in mtr((t[S^hP^i ..R^j] = t'[S^hP^i ..R^j] \land \Theta(t, \alpha, m, [1,k]) \land \Theta(t', \alpha, m, [1,k])) \rightarrow t[\dot{Y}] = t'[\dot{Y}])
\]
Computing APFDs

We compute all the APFDs, adopting a tractable sub-optimal solution and considering the three errors, $g_3$, $h_3$, $j_3$. Given a KSPE instance $w$ and the predicted attribute $\hat{B}$, our approach is mainly based on the following steps:

- Derive $s$ by TANE, such that $g_3 \leq \varepsilon_g$;
- Check on $s$ that $h_3 \leq \varepsilon_h$;
- If the previous check is fine, check $j_3 \leq \varepsilon_j$. 