EFFICIENTLY MINING CLOSED INTERVAL PATTERNS WITH CONSTRAINT PROGRAMMING

D. Bekkoucha[1], A. Ouali[1], P. Boizumault[1], B. Crémilleux[1]

[1]Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC, Caen, FRANCE

djawad.bekkoucha@unicaen.fr

Outline

- Context:
  - Mining numerical datasets
- Interval patterns
- Contributions:
  - Reified model
  - Global constraint
- Experimental results
- Conclusion and Perspectives

Data Mining

- Data mining reveals implicit relationships in a large volume of data

	Height m_1	Weight m_2	Age m_3	Severe form c
g_1	155	74	80	1
g_2	176	99	74	0
g_3	167	73	28	0
g_4	153	76	52	1
g_5	190	99	76	0

Table: Numerical dataset $\mathcal{N}$

- People with a height between [153, 155], weight between [74, 76] and age between [52, 80] are more exposed to extreme forms of a certain disease

Notation

- $\mathcal{G}$: Set of objects in $\mathcal{N}$
- $\mathcal{M}$: Set of attributes in $\mathcal{N}$
- $\mathcal{N}_m$: Set of numerical values contained in attribute $m \in \mathcal{M}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Interval Patterns

Definition

An interval pattern $\mathcal{V}$ is a vector of $|\mathcal{M}|$intervals where each interval corresponds to an attribute $m \in \mathcal{M}$

$$\mathcal{V} = \langle [a_i, b_i]_{i \in \{1,\dots,|\mathcal{M}|\}} \rangle, \; a_i, b_i \in \mathcal{N}_i \quad \wedge \quad a_i \leq b_i$$

Example

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

- $\langle [153, 190][73, 99][28, 80] \rangle$
- $\langle [153, 155][73, 76][52, 80] \rangle$
- $\langle [153, 155][74, 76][52, 80] \rangle$
- $\langle [153, 167][74, 76][52, 80] \rangle$

A set of interval patterns

Interval Patterns

Definitions

- **Cover:** $cover(\mathcal{V}) = \{g \in \mathcal{G} | \bigwedge_{m \in \mathcal{M}} \underline{x}_m \leq v_{g,m} \leq \overline{x}_m \; s.t. \; v_{g,m} \in \mathcal{N}_m\}$
- **Frequency:** $freq(\mathcal{V}) = |cover(\mathcal{V})|$
- **Description:**
  $desc(G \subseteq \mathcal{G}) = \langle [a_m, b_m] \rangle_{m \in \{1,\dots,|\mathcal{M}|\}} \; s.t. \; a_m = min(\{v_{g,m} | \, g \in G\}) \; \wedge \; b_m = max(\{v_{g,m} | \, g \in G\})$

Example

- $cover(\langle [153, 155][73, 76][52, 80] \rangle) = \{g_1, \; g_4\}$
- $freq(\langle [153, 155][73, 76][52, 80] \rangle) = |\{ \, g_1, \; g_4\}| = 2$
- $desc(\{g_1, \; g_4\}) = \langle [153, 155][74, 76][52, 80] \rangle$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Interval Patterns

Limitations

The enumeration of all the interval patterns leads to:

- Combinatorial explosion in the number of patterns
- Redundancy of the extracted interval patterns

Example

- $\langle[153,155][73,76][52,80]\rangle,\ \{g_1,g_4\}$
- $\langle[153,155][74,76][28,80]\rangle,\ \{g_1,g_4\}$
- $\langle[153,167][74,76][52,80]\rangle,\ \{g_1,g_4\}$
- $\langle[153,155][74,76][52,80]\rangle,\ \{g_1,g_4\}$

Redundant Interval Patterns

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Closed Interval Patterns

Closure

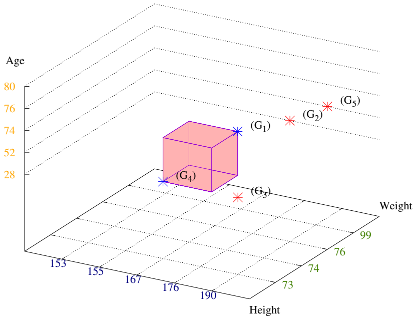
An interval pattern $\mathcal{V}$ is closed if there does not exists $\mathcal{V}'$ sharing the same support and having strictly smaller intervals than those of $\mathcal{V}$.

$$close(\mathcal{V}) \iff desc(cover(\mathcal{V})) = \mathcal{V}$$

Example for 3 attributes

- $\langle[153, 155][73, 76][52, 80]\rangle, \ \{g_1, g_4\}$
- $\langle[153, 155][74, 76][28, 80]\rangle, \ \{g_1, g_4\}$
- $\langle[153, 167][74, 76][52, 80]\rangle, \ \{g_1, g_4\}$
- $\langle[153, 155][74, 76][52, 80]\rangle, \ \{g_1, g_4\}$

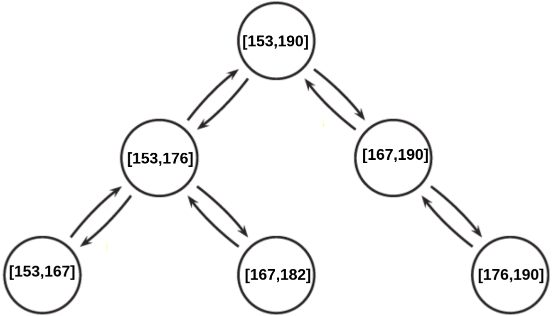
Condensed representation



Mining Closed Interval Patterns

Existing approaches

- **Dedicated approach:** [Kaytoue and al. 2011] present **MinIntChange**, a dedicated approach for mining closed interval patterns
  - Lack of genericity



Declarative Approaches for Binary data

- Itemsets: [De Raedt et al. KDD 2008], [khiari et al. CP 2010] ,[Schaus et al. CP 2017] , [Mamaar et al. CP 2016], [Belaid et al. SDM 2019]
- Sequential patterns: [Kemmar et al. Constraints 2017], [Aoga et al. ECML/PKDD 2016], [Négrevergne et al. CPAIOR 2015]
- Sky Patterns: [Ugarte et al. 2017], [Vernerey et al. IJCAI 2022], [Négrevergne et al. ICDM 2013], [Ugarte et al. ICTAI 2015]
- Top-K patterns: [Jabbour et al. ECML/PKDD 2013], [Hidouri et al. DaWaK 2021],

What about numerical Data ?

Binarization

- Binarize numerical data with **Interordinal Scaling** to avoid information loss
- Create pairs of binary attributes (items) for each numerical value:
  $\forall m \in \mathcal{M}, g \in \mathcal{G} \;\; m \leq w_{g,m}$ and $m \geq w_{g,m}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

IS binarization →

	Height				Weight	Age			
	$m_1 \leq 153$	$m_1 \geq 153$	$m_1 \leq 155$	$m_1 \geq 155$	$\cdots$	$m_3 \leq 28$	$m_3 \geq 28$	$m_3 \leq 52$	$m_3 \geq 52$
g_1	0	1	1	1	$\cdots$	0	1	0	1
g_2	0	1	0	1	$\cdots$	0	1	0	1
g_3	0	1	0	1	$\cdots$	1	1	1	1
g_4	1	1	1	0	$\cdots$	0	1	0	1
g_5	0	1	0	1	$\cdots$	0	1	0	1

Closed Interval Pattern

$\langle [153, 155][74, 76][52, 80] \rangle$

Binarization →

← **Reverse binarization**

Closed Itemset

$\{m1 \leq 153, m1 \geq 153, m1 \leq 155, m1 \geq 155, ..., m3 \leq 74, m3 \geq 76, m3 \leq 80, m3 \geq 80\}$

Contributions

Since there is no declarative approach for mining closed interval patterns, we present:

- A reified model named **CP4CIP** for mining closed interval patterns **without prior binarization**
- A global constraint named **GC4CIP** for mining closed interval patterns **without prior binarization**

First Model using Reified Constraints

Modeling intervals

Decision variables: Variables representing the borders of intervals:

$$\forall m \in \mathcal{M}, \ \ \underline{x}, \overline{x} : \ \ \mathcal{D}(\underline{x}_m) = \mathcal{D}(\overline{x}_m) = \mathcal{N}_m$$

Example

- $\mathcal{D}(\underline{x}_{m_1}) = \mathcal{D}(\overline{x}_{m_1}) = \{153, 155, 167, 176, 190\}$
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}$
- $\mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

First Model using Reified Constraints

Inclusion

Inclusion variables: $\forall m \in \mathcal{M}, g \in \mathcal{G}, B_{g,m} : \mathcal{D}(B_{g,m}) = \{0, 1\}$

- Used in the **inclusion constraints**:

$$\forall m \in \mathcal{M},\ g \in \mathcal{G},\ B_{g,m} = 1 \iff min(\mathcal{D}(\underline{x}_m)) \leq v_{g,m} \leq max(\mathcal{D}(\overline{x}_m))$$

Example

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

	Height $\in [153, 155]$	Weight $\in [74, 76]$	Age $\in [52, 80]$
g_1	1	1	1
g_2	0	0	1
g_3	0	0	0
g_4	1	1	1
g_5	0	0	1

Table: Inclusion variables for $\langle[153, 155][74, 76][52, 80]\rangle$

First Model using Reified Constraints

Coverage

Coverage variables: $\forall g \in \mathcal{G},\ y_g : \mathcal{D}(y_g) = \{0, 1\}$

- Used in **coverage constraints**

$$\forall g \in \mathcal{G}, y_g = 1 \iff \sum_{m \in \mathcal{M}} B_{g,m} = |\mathcal{M}|$$

Example

	Height $\in [153, 155]$	Weight $\in [74, 76]$	Age $\in [52, 80]$
g_1	1	1	1
g_2	0	0	1
g_3	0	0	0
g_4	1	1	1
g_5	0	0	1

Table: Inclusion variables for $\langle[153, 155][74, 76][52, 80]\rangle$

$\mathcal{D}(y_g)$	
$y_1 =$	1
$y_2 =$	0
$y_3 =$	0
$y_4 =$	1
$y_5 =$	0

Table: coverage variables

First Model using Reified Constraints

Closure

Closure variables: $\forall g \in \mathcal{G}, m \in \mathcal{M},\ \underline{H}_{g,m}, \overline{H}_{g,m} : \begin{cases} \mathcal{D}(\underline{H}_{g,m}) = \{v_{g,m}\} \cup \{\mathcal{N}_m^{\uparrow} + 1\} \\ \mathcal{D}(\overline{H}_{g,m}) = \{v_{g,m}\} \cup \{\mathcal{N}_m^{\downarrow} - 1\} \end{cases}$

- Used in **closure constraints**

$$\forall g \in \mathcal{G}, m \in \mathcal{M} \begin{cases} y_g = 1 \implies \mathcal{D}(\underline{H}_{g,m}) = \mathcal{D}(\overline{H}_{g,m}) = \{v_{g,m}\} \\ y_g = 0 \implies \mathcal{D}(\underline{H}_{g,m}) = \{\mathcal{N}_m^{\uparrow} + 1\},\ \mathcal{D}(\overline{H}_{g,m}) = \{\mathcal{N}_m^{\downarrow} - 1\} \end{cases}$$

$$\forall m \in \mathcal{M} \begin{cases} \underline{x}_m = min(\mathcal{D}(\underline{H}_{1,m}), \mathcal{D}(\underline{H}_{2,m}), ..., \mathcal{D}(\underline{H}_{|\mathcal{G}|,m})), \\ \overline{x}_m = max(\mathcal{D}(\overline{H}_{1,m}), \mathcal{D}(\overline{H}_{2,m}), ..., \mathcal{D}(\overline{H}_{|\mathcal{G}|,m})) \end{cases}$$

$\mathcal{D}(y_g)$	
$y_1 =$	1
$y_2 =$	0
$y_3 =$	0
$y_4 =$	1
$y_5 =$	0

Table: coverage variables

	$\underline{H}_{g,1}$	$\underline{H}_{g,2}$	$\underline{H}_{g,3}$
g_1	155	74	80
g_2	191	100	81
g_3	191	100	81
g_4	153	76	52
g_5	191	100	81
min:	153	74	52

	$\overline{H}_{g,1}$	$\overline{H}_{g,2}$	$\overline{H}_{g,3}$
g_1	155	74	80
g_2	152	72	27
g_3	152	72	27
g_4	153	76	52
g_5	152	72	27
max:	155	76	80

Table: Values of closure variables $\underline{H}_{g,m}$ and $\overline{H}_{g,m}$ for the running example.

First Model using Reified Constraints

Model complexity

Variables:

- Interval representation: $2 \cdot |\mathcal{M}|$
- Coverage representation: $|\mathcal{G}|$
- Closure representation: $3 \cdot |\mathcal{G}| \cdot |\mathcal{M}|$

Constraints:

- Inclusion constraints: $|\mathcal{G}| \cdot |\mathcal{M}|$
- Coverage constraints: $|\mathcal{G}|$
- Closure constraints: $4 \cdot |\mathcal{G}| \cdot |\mathcal{M}| + 2 \cdot |\mathcal{M}|$

Can we do better ?

Second Model using Global Constraints

GREYC
Electronics and Computer Science Laboratory

Why global constraints ?

- Dedicated filtering algorithm
- Captures global relations within variables
- Simplifies the problem modeling
- Preserves the genericity

$GC4CIP$

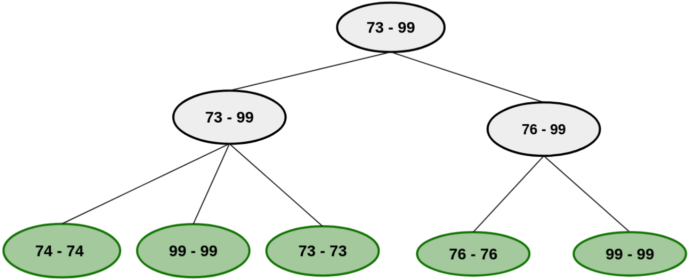
Let $\mathcal{V}$ an interval pattern. The $GC4CIP_{\mathcal{N},\theta}(\mathcal{V})$ global constraint holds iff:
- $\mathcal{V}$ is closed, and
- $\mathcal{V}$ is frequent (i.e. $freq(\mathcal{V}) \geq \theta$)

Second Model using Global Constraints

Specific data structure

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

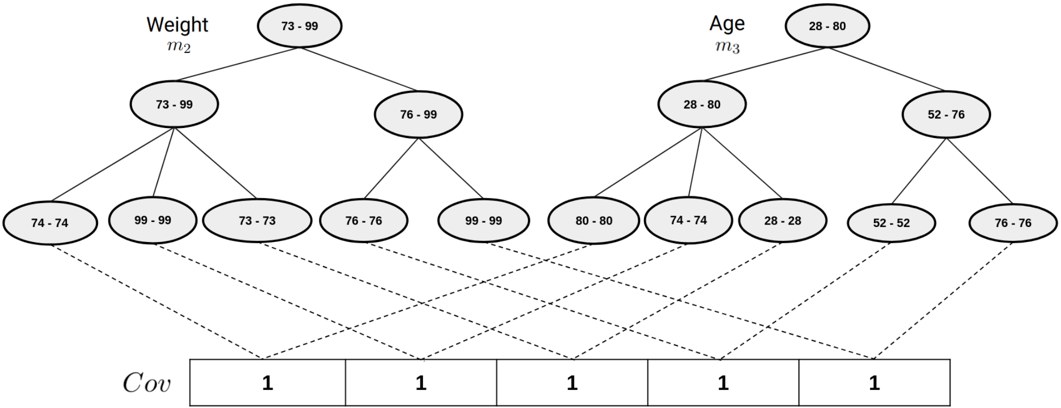


Tree corresponding to the weight attribute in $\mathcal{N}$

Second Model using Global Constraints

Specific data structure

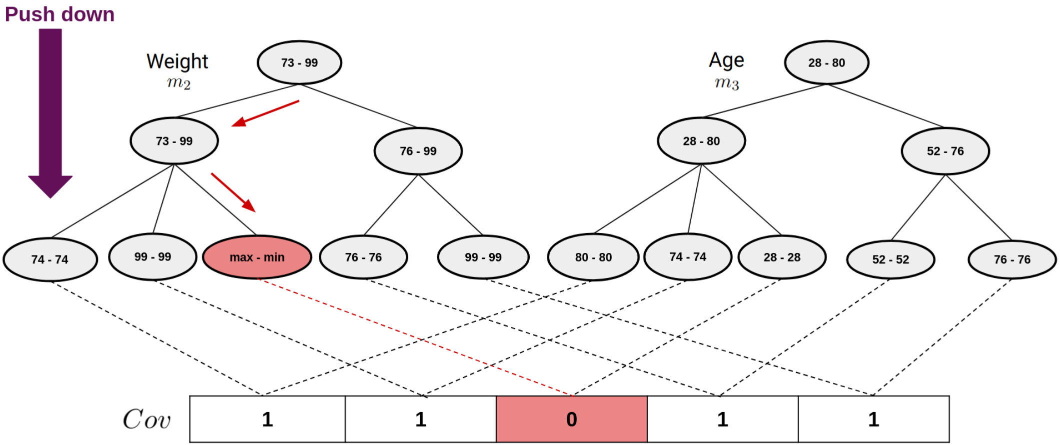
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}, \mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$



Second Model using Global Constraints

Specific data structure

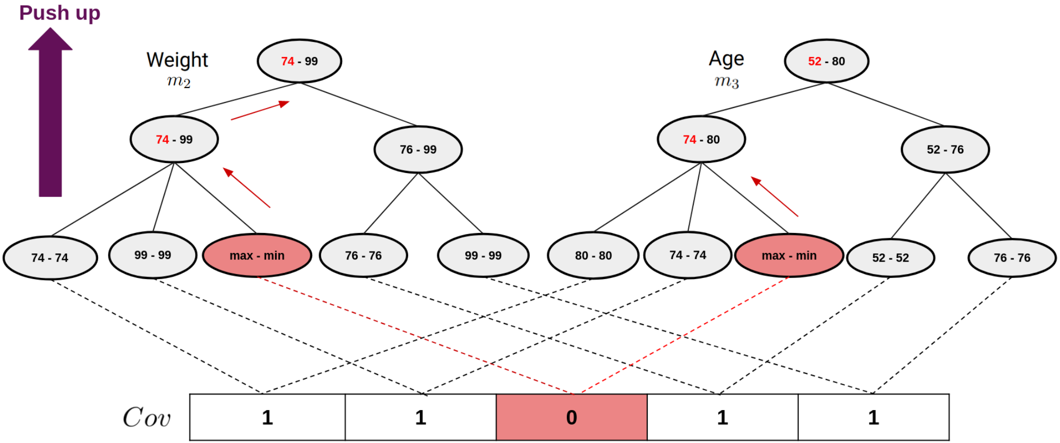
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}, \mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$



Second Model using Global Constraints

Specific data structure

- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}, \mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$



Second Model using Global Constraints

Filtering rules

Proposition 1

Let $\mathcal{V}^* = \langle[\min(\mathcal{D}(\underline{x}_1)), \max(\mathcal{D}(\overline{x}_1))], \ldots, [\min(\mathcal{D}(\underline{x}_{|\mathcal{M}|})), \max(\mathcal{D}(\overline{x}_{|\mathcal{M}|}))]\rangle$

$$\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m), \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \end{cases} \quad if: \begin{cases} \exists\, m' \in \mathcal{M}, m \neq m', v_{g,m'} < \min(\mathcal{D}(\underline{x}_{m'})) \vee v_{g,m'} > \max(\mathcal{D}(\overline{x}_{m'})) \\ \wedge \\ \forall g' \in \mathcal{G},\ g \neq g' \text{ such that } g' \text{ is covered by } \mathcal{V}^*, v_{g,m} \neq v_{g',m} \end{cases}$$

Example

During the search we have:

- $\mathcal{D}(\underline{x}_{m_1}) = \mathcal{D}(\overline{x}_{m_1}) = \{153, 155, 167, 190\}$
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}$
- $\mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Second Model using Global Constraints

Filtering rules

Proposition 1

Let $\mathcal{V}^* = \langle[\min(\mathcal{D}(\underline{x}_1)), \max(\mathcal{D}(\overline{x}_1))], \ldots, [\min(\mathcal{D}(\underline{x}_{|\mathcal{M}|})), \max(\mathcal{D}(\overline{x}_{|\mathcal{M}|}))]\rangle$

$$\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m), \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \end{cases} \quad if: \begin{cases} \exists\, m' \in \mathcal{M}, m \neq m', v_{g,m'} < \min(\mathcal{D}(\underline{x}_{m'})) \vee v_{g,m'} > \max(\mathcal{D}(\overline{x}_{m'})) \\ \wedge \\ \forall g' \in \mathcal{G},\ g \neq g' \text{ such that } g' \text{ is covered by } \mathcal{V}^*, v_{g,m} \neq v_{g',m} \end{cases}$$

Example

During the search we have:

- $\mathcal{D}(\underline{x}_{m_1}) = \mathcal{D}(\overline{x}_{m_1}) = \{153, 155, 167, 190\}$
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}$
- $\mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, \mathbf{74}, 76, 80\}$

Removing 74 from $\mathcal{D}(\underline{x}_{m_3})$ and $\mathcal{D}(\overline{x}_{m_3})$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	**74**
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Second Model using Global Constraints

Filtering rules

Proposition 2

Let $m, m' \in \mathcal{M}$, $m \neq m'$ $\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m) \text{ if: } & v_{g,m} > max(join(x_{m'}, \underline{x}_m)) \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \text{ if: } & v_{g,m} < min(join(x_{m'}, \overline{x}_m)) \end{cases}$

Example

During the search the domain of $\mathcal{D}(\overline{x}_2)$ has changed. We have:

	Height m_1		Weight $m2$		Age m_3	
domains	$\mathcal{D}(\underline{x}_1)$	$\mathcal{D}(\overline{x}_1)$	$\mathcal{D}(\underline{x}_2)$	$\mathcal{D}(\overline{x}_2)$	$\mathcal{D}(\underline{x}_3)$	$\mathcal{D}(\overline{x}_3)$
g_1	155	155	74	74	80	80
g_2	176	176	99	99	74	74
g_3	167	167	73	73	28	28
g_4	153	153	76	76	52	52
g_5	190	190	99	99	76	76

Table: Domains of the attributes variables

Second Model using Global Constraints

Filtering rules

Proposition 2

Let $m, m' \in \mathcal{M}$, $m \neq m'$ $\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m) \text{ if: } & v_{g,m} > max(join(x_{m'}, \underline{x}_m)) \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \text{ if: } & v_{g,m} < min(join(x_{m'}, \overline{x}_m)) \end{cases}$

Example

During the search the domain of $\mathcal{D}(\overline{x}_2)$ has changed. We have:

- Propagate the partial domain knowledge on $\overline{x}_2$ to other domains

	Height m_1		Weight $m2$		Age m_3	
domains	$\mathcal{D}(\underline{x}_1)$	$\mathcal{D}(\overline{x}_1)$	$\mathcal{D}(\underline{x}_2)$	$\mathcal{D}(\overline{x}_2)$	$\mathcal{D}(\underline{x}_3)$	$\mathcal{D}(\overline{x}_3)$
g_1	155	155	74	74	80	80
g_2	176	176	99	99	74	74
g_3	167	167	73	73	28	28
g_4	153	153	76	76	52	52
g_5	190	190	99	99	76	76

Table: Domains of the attributes variables

Second Model using Global Constraints

Filtering rules

Proposition 2

Let $m, m' \in \mathcal{M}$, $m \neq m'$ $\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m) \text{ if: } & v_{g,m} > max(join(x_{m'}, \underline{x}_m)) \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \text{ if: } & v_{g,m} < min(join(x_{m'}, \overline{x}_m)) \end{cases}$

Example

During the search the domain of $\mathcal{D}(\overline{x}_2)$ has changed. We have:

- Propagate the partial domain knowledge on $\overline{x}_2$ to other domains
- $join(\overline{x}_2, \overline{x}_3) = join(\overline{x}_2, \underline{x}_3) = \{28, 52, 76\}$

	Height m_1		Weight $m2$		Age m_3	
domains	$\mathcal{D}(\underline{x}_1)$	$\mathcal{D}(\overline{x}_1)$	$\mathcal{D}(\underline{x}_2)$	$\mathcal{D}(\overline{x}_2)$	$\mathcal{D}(\underline{x}_3)$	$\mathcal{D}(\overline{x}_3)$
g_1	155	155	74	74	80	80
g_2	176	176	99	99	74	74
g_3	167	167	73	73	28	28
g_4	153	153	76	76	52	52
g_5	190	190	99	99	76	76

Table: Domains of the attributes variables

Second Model using Global Constraints

Filtering rules

Proposition 2

Let $m, m' \in \mathcal{M}, m \neq m'$ $\begin{cases} v_{g,m} \notin \mathcal{D}(\underline{x}_m) \text{ if: } & v_{g,m} > max(join(x_{m'}, \underline{x}_m)) \\ v_{g,m} \notin \mathcal{D}(\overline{x}_m) \text{ if: } & v_{g,m} < min(join(x_{m'}, \overline{x}_m)) \end{cases}$

Example

During the search the domain of $\mathcal{D}(\overline{x}_2)$ has changed. We have:

- Propagate the partial domain knowledge on $\overline{x}_2$ to other domains
- $join(\overline{x}_2, \overline{x}_3) = join(\overline{x}_2, \underline{x}_3) = \{28, 52, 76\}$
- $80 > max(join(\overline{x}_2, \underline{x}_3))$ then remove 80 from $\mathcal{D}(\underline{x}_3)$

	Height m_1		Weight $m2$		Age m_3	
domains	$\mathcal{D}(\underline{x}_1)$	$\mathcal{D}(\overline{x}_1)$	$\mathcal{D}(\underline{x}_2)$	$\mathcal{D}(\overline{x}_2)$	$\mathcal{D}(\underline{x}_3)$	$\mathcal{D}(\overline{x}_3)$
g_1	155	155	74	74	80	80
g_2	176	176	99	99	74	74
g_3	167	167	73	73	28	28
g_4	153	153	76	76	52	52
g_5	190	190	99	99	76	76

Table: Domains of the attributes variables

Second Model using Global Constraints

Filtering rules

Proposition 3

Let $m \in \mathcal{M}$, and $\mathcal{V}^p = \langle[\min(\mathcal{D}(\underline{x}_i)), \max(\mathcal{D}(\overline{x}_i))]$

- $a_m \notin \mathcal{D}(\underline{x}_m)$ if $freq(\mathcal{V}^p \quad ++ \quad [a_m, \max(\mathcal{D}(\overline{x}_m))]) < \theta$
- $b_m \notin \mathcal{D}(\overline{x}_m)$ if $freq(\mathcal{V}^p \quad ++ \quad [\min(\mathcal{D}(\underline{x}_m)), b_m]) < \theta$

Example

Let $\theta = 2$ and suppose the following variables domains:

- $\mathcal{D}(\underline{x}_{m_1}) = \{176, 190\},\ \mathcal{D}(\overline{x}_{m_1}) = \{176, 190\}$
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}$
- $\mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

Second Model using Global Constraints

Filtering rules

Proposition 3

Let $m \in \mathcal{M}$, and $\mathcal{V}^p = \langle[\min(\mathcal{D}(\underline{x}_i)), \max(\mathcal{D}(\overline{x}_i))]$

- $a_m \notin \mathcal{D}(\underline{x}_m)$ if $freq(\mathcal{V}^p \quad ++ \quad [a_m, \max(\mathcal{D}(\overline{x}_m))]) < \theta$
- $b_m \notin \mathcal{D}(\overline{x}_m)$ if $freq(\mathcal{V}^p \quad ++ \quad [\min(\mathcal{D}(\underline{x}_m)), b_m]) < \theta$

Example

Let $\theta = 2$ and suppose the following variables domains:

- $\mathcal{D}(\underline{x}_{m_1}) = \{176, \mathbf{190}\}$, $\mathcal{D}(\overline{x}_{m_1}) = \{176, 190\}$
- $\mathcal{D}(\underline{x}_{m_2}) = \mathcal{D}(\overline{x}_{m_2}) = \{73, 74, 76, 99\}$
- $\mathcal{D}(\underline{x}_{m_3}) = \mathcal{D}(\overline{x}_{m_3}) = \{28, 52, 74, 76, 80\}$

	Height m_1	Weight m_2	Age m_3
g_1	155	74	80
g_2	176	99	74
g_3	167	73	28
g_4	153	76	52
g_5	190	99	76

Table: Numerical dataset $\mathcal{N}$

- $freq(\langle[176, max(\overline{x}_m)] + +\mathcal{V}^p\rangle) = 2 \geq \theta$ then 176 is maintained in $\mathcal{D}(\underline{x}_{m_1})$
- $freq(\langle[\mathbf{190}, max(\overline{x}_m)] + +\mathcal{V}^p\rangle) = 1 < \theta$ then **Filter** 190 from $\mathcal{D}(\underline{x}_{m_1})$

Second Model using Global Constraints

Model complexity

- The push down and push up has a worst case complexity of $\mathcal{O}(|\mathcal{G}|)$. This is simplified from $O(S^{log_S|\mathcal{G}|})$, where S is the maximal number of children of a parent node.
- The GC4CIP worst case complexity is $\mathcal{O}(|\mathcal{M}| \cdot |\mathcal{G}|^3 \cdot log_S|\mathcal{G}|)$

Experimental protocol

Configuration:

- ORTools CP-Solver version 9.0 (C++)
- 5 hours timeout
- 512 GB of memory limit

Benchmark of numerical datasets:

	NT	AP	BK	Cancer	CH	Yacht	LW
$\lvert\mathcal{M}\rvert$	3	5	5	9	8	7	10
$\lvert\mathcal{G}\rvert$	130	135	96	116	209	308	189
#distinct values	67	674	313	900	396	322	253
	Interordinal scaled datasets						
$\#$Binary attributes	134	1348	626	1800	792	644	506

Experimental protocol

Compared approaches

We compared our approaches **CP4CIP** and **GC4CIP** to:

- **Dedicated approaches**
  - **MinIntChange:** a closed interval pattern mining approach that does not require any pre-post processing step
- **Declarative approaches**
  - **CP4IM:** a reified model for mining closed patterns (itemsets) from binary data.
  - **CLOSEDPATTERN:** a global constraint for mining closed patterns (itemsets) from binary data.

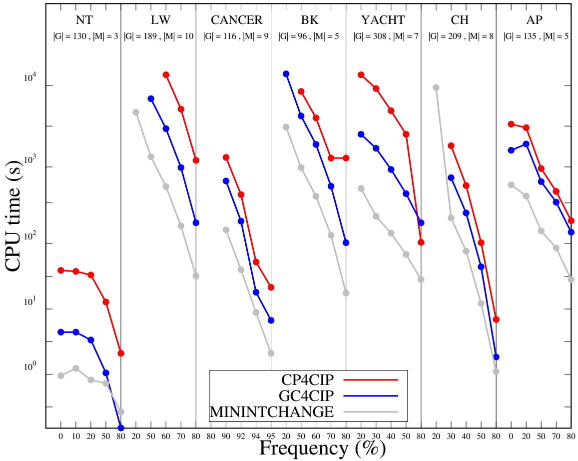
Note: The comparison with **CP4IM** and **CP4CIP** requires pre-processing and post-processing steps to handle numerical data.

Experimental results

- CP4CIP and GC4CIP have better scalability then other approaches
- CP4CIP outperforms CP4IM in most of instances
- GC4CIP outperforms CLOSEDPATTERN in all instances

$\mathcal{N}$	θ (%)	# Sol ($\approx$)	CPU Time (s) CP4IM	CLOSEDPATTERN	p-p-processing	CP4IM+p-p-p	CLOSEDPATTERN+p-p-p	CP4CIP	GC4CIP
BK	80	10^6	1840.21	148.91	176.65	2016.86	325.56	**271.10**	89.63
	70	10^7	15132.87	1457.99	1326.58	16459.45	2784.57	**1770.22**	655.63
	60	10^7	TO	8643.34	6713.25	TO	15356.59	**7311.24**	2879.54
	50	10^8	TO	28302.62	19307.70	TO	47610.32	**18471.23**	7780.65
	20	10^8	TO	TO	TO	TO	TO	TO	34598.10
Cancer	95	10^4	170.14	6.19	13.69	183.83	19.88	**18.42**	5.80
	94	10^5	568.00	18.21	38.88	606.88	57.09	**45.43**	15.66
	92	10^5	6944.07	294.14	542.82	7486.89	836.96	**486.87**	190.84
	90	10^6	29787.19	1190.42	2348.45	32135.64	3538.87	**1806.19**	786.25
AP	80	10^5	783.92	175.02	55.21	839.13	230.23	**28.55**	19.18
	70	10^6	5909.86	189.30	415.76	6325.62	605.06	**194.64**	128.83
	60	10^6	18479.87	7995.84	1275.85	19755.72	9271.69	**548.12**	373.01
	50	10^7	TO	23252.89	2964.71	TO	26217.60	**1223.79**	770.83
	20	10^7	TO	43199.73	3052.93	TO	46252.66	**5129.20**	2891.55
	0	10^7	TO	TO	TO	TO	TO	**5867.37**	2343.98
CH	95	10^6	25.59	1.16	29.93	55.52	31.09	**5.98**	1.60
	90	10^5	608.94	36.58	224.70	833.64	261.28	**89.81**	38.42
	85	10^6	4753.35	331.08	835.24	5588.59	1166.32	**671.49**	256.86
	80	10^6	19154.96	1444.64	18009.40	37164.36	19454.04	**2739.85**	890.82
	50	TO	TO	TO	TO	TO	TO	TO	TO
LW	80	10^6	1612.68	96.91	174.46	1787.14	271.37	**1638.03**	181.81
	70	10^6	12904.12	757.02	1279.34	14183.34	2036.36	**9886.90**	1269.50
	60	10^7	TO	3436.91	5236.91	TO	8673.82	**33 148.24**	4,965.20
	50	10^8	TO	11060.23	15588.10	TO	26648.33	TO	14298.64
	20	TO	TO	TO	TO	TO	TO	TO	TO
NT	80	10^3	0.87	0.06	0.07	**0.97**	0.13	1.80	0.13
	50	10^4	7.08	0.41	0.50	**7.58**	0.91	11.01	0.91
	20	10^4	28.13	1.53	1.83	29.96	3.36	**28.77**	2.89
	10	10^5	41.75	2.51	2.61	44.36	5.12	**32.50**	4.02
	0	10^5	62.48	2.88	3.13	65.61	6.01	**33.72**	3.81
Yacht	80	10^4	40.12	2.03	83.20	123.32	85.23	**90.92**	2.45
	50	10^6	7277.85	336.03	268.28	7546.13	604.31	**4090.63**	181.63
	40	10^6	30519.66	1282.32	727.09	31246.75	2009.41	**9380.16**	501.52
	30	10^7	TO	4265.71	1695.63	TO	5961.34	**20464.22**	1179.13
	20	10^7	TO	12898.20	2874.08	TO	15772.28	**33294.36**	2487.68
	0	10^7	TO	TO	TO	TO	TO	TO	4116.60

Experimental results



Conclusion

- We presented two declarative approaches for mining closed interval patterns:
  - A reified model denoted **CP4CIP**
  - A global constraint denoted **GC4CIP**

- We demonstrated the efficiency of mining interval patterns directly from numerical data

Perspectives

- Improve the filtering algorithm of GC4CIP with a different data structure
- Reduce the amount of mined Interval Patterns by:
  - mining diversified interval patterns
  - mining patterns according to a user feedback (interactive pattern mining)

The end

Thank you

Any Questions ?

djawad.bekkoucha@unicaen.fr