Constraint Programming for Itemset Mining

Tias Guns, K.U.Leuven, Belgium

In collaboration with Siegfried Nijssen and Luc De Raedt

Based on papers at KDD08 and KDD09
Position in summer school

Itemset Mining (Bart Goethals' talk)
- Apriori (Level-wise search, anti-monotonicity)
- Eclat (Specific depth-first search)

Constraint Programming
- Combinatorial Satisfaction Problems (CSP)
- Generic depth-first search
I. Motivation, constraint-based mining
II. Constraint Programming basics
III. Constraint-based itemset mining using CP
IV. Correlated itemset mining using CP
V. Conclusions.
(frequent) Itemset mining

Transactions:

1) {Doritos, Fritos}

2) {Pampers, Playtex}

3) {beer, onion, beer}

4) {Fritos, Pampers, Playtex}

5) {beer, onion, beer}

6) {Doritos, Fritos, Pampers, Playtex}

Patterns:

7) {Doritos, Fritos, Pampers, Playtex}

8) {beer, onion, beer}

9) {Fritos, Pampers, Playtex}

10) {beer, onion, beer}

11) {beer, onion, beer}

12) {beer, onion, beer}

{Doritos, Fritos} (42%)

{beer, onion, beer} (33%)
Goal: find patterns in transactional data
- better understanding of data
- find novel information

Solution: Itemset Mining

Applications:
- online shops
- weblog analysis
- microarray analysis (gene expression)
- learning taxonomies
- text analysis (privacy leaks)
- ...
(frequent) Itemset mining

Transactions:

1) {Doritos, Coca-Cola, Pringles} (83%)
2) {McDonald's} (58%)
3) {Doritos} (33%)
4) {Pampers} (42%)
5) {Doritos} (33%)
6) {Doritos, Pringles, Heineken} (50%)
7) {Pampers} (42%)
8) {Doritos, Coca-Cola} (50%)
9) {Doritos, Pampers} (42%)
10) {Doritos, Heineken} (33%)
11) {Doritos, Heineken} (42%)
12) {Doritos, Pampers} (42%)
- Time-consuming to interpret
- Long algorithmic runtime
Goal: find patterns in transactional data

Solution: Itemset Mining

Problem: too many patterns

Solution: Constraint-based Itemset Mining
    select only interesting patterns, based on domain knowledge
Use of constraints in data mining to specify the desired set of solutions (Mannila & Toivonen, 97)

\[ Th(\mathcal{L}, Q, \mathcal{D}) = \{ p \in \mathcal{L} | Q(p, \mathcal{D}) = \text{true} \} \]

- \( \mathcal{L} = 2^I \), i.e., itemsets
- \( \mathcal{D} \subset \mathcal{L} \), i.e., transactions
- \( Q(p, \mathcal{D}) = \text{true} \) if \( \text{freq}(p, \mathcal{D}) \geq t \)
**Constraint-based Itemset Mining**

- condensed representations
  - Maximal patterns: remove all redundancy
  - Closed patterns: remove redundancy, keep frequencies
  - \textit{delta}-closed patterns: closed + fault tolerance
- user defined constraints
  - human readable \( \rightarrow \) \( \text{size}(\text{itemset}) \leq 5 \)
  - high value \( \rightarrow \) \( \text{total\_cost}(\text{itemset}) \geq 100 \) £
  - infrequent on other dataset \( \rightarrow \) \( \text{freq\_part2}(\text{itemset}) \leq 1\% \)

...
+ many constraints proposed
- new constraint often require new implementations
- combining constraints?

state-of-the-art = hard-coded support for some popular constraint families.

=> No principled approach
The need for a principled approach

The Data Mining process model:
I. Motivation, constraint-based mining

II. Constraint Programming basics

III. Constraint-based itemset mining using CP

IV. Correlated itemset mining using CP

V. Conclusions.
Constraint programming:

- ... solves combinatorial satisfaction problems
- ... is used in many applications
- ... is an active research area
- ... is among the most efficient general problem solving techniques
How CP works

Constraint Programming =

MODEL (by user) +

SEARCH (by solver)
A CP model

- **variables**
  \[ [E_{11} \ldots E_{99}] \]

- **domains**
  \[ E_{xy} = \{1 \ldots 9\} \]

- **constraints**
  allDifferent([E_{1x}]), ...
  allDifferent([E_{xy}]), ...
  allDifferent([E_{11} \ldots E_{33}]), ...
Two key principles:

- **Propagation** of constraints
  
  eg. `alldiff(X,Y,Z) X={1}, Y={1,2}, Z={1,2,3,4} → Y={2}, Z={3,4}`

  Every constraint is implemented by a propagator.

- **Branch** over values of variables
  
  eg. Propagation at fixpoint → branch over `Z={3}`

  Search is recursive and complete
A CP search

all rows: all_different(row)
all columns: all_different(col)
all squares: all_different(square)

CP: Branch & Propagate

- propagate 2 (row)
- branch 4
- propagate 6 (square)
I. Motivation, pattern mining
II. Constraint Programming basics
III. Constraint-based itemset mining using CP
IV. Correlated itemset mining using CP
V. More pattern mining at work
VI. Conclusion.
Surprisingly, Constraint Programming had not been used for constraint-based mining yet...

Constraint Programming for Itemset Mining
in short: (KDD2008)

- using out-of-the-box CP solvers
- allows to express many IM constraints
- easily combine all those constraints
# Itemset mining

## Transactions:

<table>
<thead>
<tr>
<th></th>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
<th>Item 5</th>
<th>Item 6</th>
<th>Item 7</th>
<th>Item 8</th>
<th>Item 9</th>
<th>Item 10</th>
<th>Item 11</th>
<th>Item 12</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>11</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>12</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>
variables
\[ [I_1 \ldots I_n], [T_1 \ldots T_m] \]

domains
\[ I_x, T_y = \{0, 1\} \]

constraints

frequency: \[ \sum_{t \in I} T_t \geq \theta. \]
- **variables**
  \[ [I_1 \ldots I_n], [T_1 \ldots T_m] \]

- **domains**
  \[ I_x, T_y = \{0, 1\} \]

- **constraints**
  
  **frequency:**
  \[ \sum_{t \in T} T_t \geq \theta. \]

  **OR freq. reified:**
  \[ \forall i \in \mathcal{I} : I_i = 1 \implies \sum_{t \in T} T_t D_{ti} \geq \theta. \]
variables
\([I_1 \ldots I_n], [T_1 \ldots T_m]\)

domains
\(I_x, T_y = \{0, 1\}\)

constraints

- frequency:
  \[\sum_{t \in T} T_t \geq \theta.\]

- OR freq. reified:
  \[\forall i \in \mathcal{I} : I_i = 1 \rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta.\]

- coverage:
  \[\forall t \in \mathcal{T} : T_t = 1 \iff \sum_{i \in \mathcal{I}} I_i (1 - D_{ti}) = 0.\]
Algorithm 1 Fim_cp’s frequent itemset mining model, in Essence’

1. given NrT, NrI : int
2. given TDB : matrix indexed by [int(1..NrT),int(1..NrI)] of int
3. given Freq : int
4. find Items : matrix indexed by [int(1..NrI)] of bool
5. find Trans : matrix indexed by [int(1..NrT)] of bool

6. such that

7. $\forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0$:

8. forall t: int(1..NrT).

9. $Trans[t] <=> ((\sum i: int(1..NrI). Items[i]*(1-TDB[t,i])) = 0),$

10. $\forall i \in I : I_i = 1 \rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta$:

11. forall i: int(1..NrI).

12. $Items[i] => ((\sum t: int(1..NrT). Trans[t]*TDB[t,i]) >= Freq)$
The FIM_CP search

coverage: \( \forall t \in T : T_t = 1 \iff \sum_{i \in \mathcal{I}} I_i (1 - D_{ti}) = 0. \)

freq \( \geq 2: \) \( \forall i \in \mathcal{I} : I_i = 1 \rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta. \)

CP: Branch & Propagate

- propagate \( i_2 \) (freq)

Intuition: infrequent

\( i_2 \) can never be part of freq. superset
The FIM_CP search

coverage: \( \forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0. \)

freq \( \geq 2 \): \( \forall i \in I : I_i = 1 \rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta. \)

CP: Branch & Propagate

- propagate i2 (freq)
- propagate t1 (coverage)

Intuition: unavoidable

\( t1 \) will always be covered
The FIM\_CP search

coverage: $\forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0.$

freq $\geq$ 2: $\forall i \in I : I_i = 1 \rightarrow \sum_{t \in T} T_tD_{ti} \geq \theta.$

CP: Branch & Propagate

- propagate $i_2$ (freq)
- propagate $t_1$ (coverage)
The FIM_CP search

coverage: \[ \forall t \in T : T_t = 1 \iff \sum_{i \in \mathcal{I}} I_i (1 - D_{ti}) = 0. \]

freq >= 2: \[ \forall i \in \mathcal{I} : I_i = 1 \implies \sum_{t \in T} T_t D_{ti} \geq \theta. \]

CP: Branch & Propagate

- propagate i2 (freq)
- propagate t1 (coverage)
- branch i1=1
- propagate t3 (coverage)

Intuition: obsolete

\[ t3 \text{ is missing an item of the itemset} \]
The FIM_CP search

coverage: \( \forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0. \)

freq >= 2: \( \forall i \in I : I_i = 1 \implies \sum_{t \in T} T_t D_{ti} \geq \theta. \)

CP: Branch & Propagate

- propagate i2 (freq)
- propagate t1 (coverage)
- branch i1=1
- propagate t3 (coverage)
- propagate i3 (freq)

Intuition: infrequent

*i3 can never be part of freq. superset*
The FIM_CP search

coverage: \( \forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0. \)

freq \( \geq 2 \): \( \forall i \in I : I_i = 1 \rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta. \)

**CP: Branch & Propagate**

- propagate i2 (freq)
- propagate t1 (coverage)
- branch i1=1
- propagate t3 (coverage)
- propagate i3 (freq)
- propagate t2 (coverage)
The FIM\_CP search

coverage:  \( \forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0. \)

freq \( \geq 2 \):  \( \forall i \in I : I_i = 1 \Rightarrow \sum_{t \in T} T_t D_{ti} \geq \theta. \)

CP: Branch & Propagate

- propagate i2 (freq)
- propagate t1 (coverage)
- branch i1=1
- propagate t3 (coverage)
- propagate i3 (freq)
- propagate t2 (coverage)
- ...

```
   t1 1 1 0 1 1
   t2 1 1 1 0 1
   t3 0 0 0 1 1
```

```
   i1  i2  i3  i4
   1  0  0  0/1
```
FIM_CP model: expressive

- **Base model (Frequent Itemset Mining)**
  \[ T_t = 1 \iff \sum_i I_i (1 - D_{ti}) = 0 \]
  \[ I_i = 1 \implies \sum_t T_t D_{ti} \geq \text{Freq} \]

- **Maximal Frequent Itemset Mining**
  \[ T_t = 1 \iff \sum_i I_i (1 - D_{ti}) = 0 \]
  \[ I_i = 1 \iff \sum_t T_t D_{ti} \geq \text{Freq} \]

- **Closed Itemset Mining**
  \[ T_t = 1 \iff \sum_i I_i (1 - D_{ti}) = 0 \]
  \[ I_i = 1 \implies \sum_t T_t D_{ti} \geq \text{Freq} \]
  \[ I_i = 1 \iff \sum_t T_t (1 - D_{ti}) = 0 \]

- **δ-Closed Itemset Mining**
  \[ T_t = 1 \iff \sum_i I_i (1 - D_{ti}) = 0 \]
  \[ I_i = 1 \implies \sum_t T_t D_{ti} \geq \text{Freq} \]
  \[ I_i = 1 \iff \sum_t T_t (1 - \delta - D_{ti}) = 0 \]
FIM_CP model: general

Course on data
Minimum frequency
Maximum frequency
Emerging patterns
Condensed Representations
Maximal
Closed
δ—Closed
Constraints on syntax
Max/Min total cost
Minimum average cost
Max/Min size

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Constraints on data</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td></td>
<td></td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Emerging patterns</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Condensed Representations</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Maximal</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Closed</td>
<td>X</td>
<td>X</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>δ—Closed</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Constraints on syntax</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max/Min total cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Minimum average cost</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Max/Min size</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

Table 1: Comparison of Itemset Miners

=> most general system to date!
FIM_CP model: flexible

combining constraints is the core of CP

=> most flexible system to date!
In Short: FIM_CP

- Principled approach
- Using generic Constraint Programming
- Declarative language, very expressive
Runtime behavior, unconstrained

Dataset properties:

<table>
<thead>
<tr>
<th>Dataset</th>
<th># items</th>
<th># transactions</th>
<th>sparseness</th>
</tr>
</thead>
<tbody>
<tr>
<td>german-credit</td>
<td>77</td>
<td>1000</td>
<td>0.28</td>
</tr>
<tr>
<td>mushroom</td>
<td>116</td>
<td>8124</td>
<td>0.17</td>
</tr>
<tr>
<td>letter</td>
<td>74</td>
<td>20000</td>
<td>0.33</td>
</tr>
</tbody>
</table>
Runtime behavior, constrained

Dataset: segment 61x2310 (sparseness: 0.51)

# patterns with min. freq. of 10% only: > 64 million
Impossible to mine unconstrained with lower freq. treshold.
Experiment conclusions

bad for

- very large datasets (> 1,000,000 transactions)
- very low frequency unconstrained (< 0.1 %)

ideal for

- studying existing constraints
- rapid prototyping of new constraints
- exploratory constraint-based mining
I. Motivation, pattern mining
II. Constraint Programming basics
III. Constraint-based itemset mining using CP
IV. Correlated itemset mining using CP
V. Conclusions.
**Correlated Itemset Mining**

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>Owns_real_estat</th>
<th>Has_savings</th>
<th>Has_loans</th>
<th>Good_customer</th>
</tr>
</thead>
<tbody>
<tr>
<td>TP: p</td>
<td>+ + + + + + +</td>
<td>- - - - + +</td>
<td>+ + + +</td>
<td>+ + + + + + +</td>
</tr>
<tr>
<td>FP: n</td>
<td>- - - - - - -</td>
<td>+ + + + + +</td>
<td>- - - -</td>
<td>- - - - - - -</td>
</tr>
<tr>
<td>FN:</td>
<td>- - - - - - -</td>
<td>- - - - - -</td>
<td>+ + + +</td>
<td>- - - - - - -</td>
</tr>
<tr>
<td>TN:</td>
<td>+ + + + + + +</td>
<td>+ + + + + +</td>
<td>+ + + +</td>
<td>+ + + + + + +</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>TP: 3 (=p)</th>
<th>FP: 0 (=n)</th>
<th>FN: 1</th>
<th>TN: 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>P:</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>N:</td>
<td>4</td>
<td>3</td>
<td>4</td>
<td>3</td>
</tr>
</tbody>
</table>
Constraint-based mining

- Frequent itemset mining (association rule mining)

  - Traditional pattern mining:
    \[ Th(\mathcal{L}, Q, \mathcal{D}) = \{ p \in \mathcal{L} | Q(p, \mathcal{D}) = true \} \]

- Correlated itemset mining (correlation rule mining)

  - Correlated pattern mining with function \( \phi(p, \mathcal{D}) \), \( (\chi^2) \):
    \[ Th(\mathcal{L}, Q, \mathcal{D}) = \arg_{p \in \mathcal{L}} \max_k \phi(p, \mathcal{D}) \]
Correlated itemset mining

Also known as:

- Discriminative itemset mining
- Contrast set mining
- Emerging itemsets
- Subgroup discovery
- Interesting itemsets

They all find an itemset/rule in labeled data that optimises a convex (correlation) measure.
ROC analysis: PN-space

Contingency Table

<table>
<thead>
<tr>
<th></th>
<th>TP: 3 (=p)</th>
<th>FP: 0 (=n)</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>FN: 1</td>
<td></td>
<td></td>
<td>3</td>
</tr>
<tr>
<td>TN: 3</td>
<td></td>
<td></td>
<td>4</td>
</tr>
<tr>
<td>P: 4</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>N: 3</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Best itemset

n

p
Measuring correlation

Many correlation functions (chi2, fisher, inf. gain) are convex and zero on the diagonal
Convex measures in CP

- **Frequent itemset mining:**
  - coverage:
    \[ \forall t \in \mathcal{T}: T_t = 1 \iff \sum_{i \in \mathcal{I}} I_i (1 - D_{ti}) = 0. \]
  - frequency:
    \[ \forall i \in \mathcal{I}: I_i = 1 \implies \sum_{t \in \mathcal{T}} T_t D_{ti} \geq \theta. \]

- **Correlated itemset mining:**
  - coverage:
    \[ \forall t \in \mathcal{T}: T_t = 1 \iff \sum_{i \in \mathcal{I}} I_i (1 - D_{ti}) = 0. \]
  - correlation:
    \[ \forall i \in \mathcal{I}: I_i = 1 \implies f\left( \sum_{t \in \mathcal{T}^+} T_t D_{ti}, \sum_{t \in \mathcal{T}^-} T_t D_{ti} \right) \geq \theta \]
  + branch and bound search
Bound in PN-space

General to specific search

- Adding an item will give equal or lower $p$ and $n$

Morishita & Sese, 2000
Improved bound in PN-space

Key observation: unavoidable transactions

<table>
<thead>
<tr>
<th>i1</th>
<th>i2</th>
<th>i3</th>
<th>i4</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/1</td>
<td>0</td>
<td>0/1</td>
<td>0/1</td>
</tr>
<tr>
<td>t1 0/1</td>
<td>1 0 1 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t2 0/1</td>
<td>1 1 0 1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>t3 0/1</td>
<td>0 0 1 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Better bound in PN-space

Key observation: unavoidable transactions

<table>
<thead>
<tr>
<th>t1</th>
<th>t2</th>
<th>t3</th>
</tr>
</thead>
<tbody>
<tr>
<td>0/1</td>
<td>0/1</td>
<td>0/1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0/1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>
Branch and propagate CIMCP

coverage:
\[ \forall t \in T : T_t = 1 \iff \sum_{i \in I} I_i (1 - D_{ti}) = 0. \]

correlation:
\[ \forall i \in I : I_i = 1 \Rightarrow f \left( \sum_{t \in T^+} T_t D_{ti}, \sum_{t \in T^-} T_t D_{ti} \right) \geq \theta \]

iterative pruning loop:
Taking the *unavoidable* transactions into account, results in more effective pruning...

**Correlated Itemset Mining in ROC space:**
A Constraint Programming Approach

in short: (KDD2009)

- based on principles of **ROC analysis**
- using insights from **Constraint Programming**
- very **fast and effective pruning**
Experiments

- Branch and bound search for top-1 pattern

- In CP:
  - 1-support (traditional minimum support)
  - 2-support (Morishita & Sese, 2000)
  - 4-support (with unavoidable transactions)
# Experiments in CP

Runtime in seconds, >900s indicated by >

<table>
<thead>
<tr>
<th>Name</th>
<th>Density</th>
<th>4-supp.</th>
<th>2-supp.</th>
<th>1-supp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>0.45</td>
<td>0.22</td>
<td>24.09</td>
<td>72.71</td>
</tr>
<tr>
<td>australian-credit</td>
<td>0.41</td>
<td>0.30</td>
<td>0.63</td>
<td>17.52</td>
</tr>
<tr>
<td>breast-wisconsin</td>
<td>0.5</td>
<td>0.28</td>
<td>13.66</td>
<td>228.08</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.5</td>
<td>2.45</td>
<td>128.04</td>
<td>&gt;</td>
</tr>
<tr>
<td>german-credit</td>
<td>0.34</td>
<td>2.39</td>
<td>66.79</td>
<td>&gt;</td>
</tr>
<tr>
<td>heart-cleveland</td>
<td>0.47</td>
<td>0.19</td>
<td>2.15</td>
<td>29.58</td>
</tr>
<tr>
<td>hypothyroid</td>
<td>0.49</td>
<td>0.71</td>
<td>10.91</td>
<td>&gt;</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.5</td>
<td>1.44</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>0.49</td>
<td>0.92</td>
<td>46.20</td>
<td>713.35</td>
</tr>
<tr>
<td>letter</td>
<td>0.5</td>
<td>52.66</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>mushroom</td>
<td>0.18</td>
<td>14.11</td>
<td>13.48</td>
<td>27.31</td>
</tr>
<tr>
<td>pendigits</td>
<td>0.5</td>
<td>3.68</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>primary-tumor</td>
<td>0.48</td>
<td>0.03</td>
<td>0.13</td>
<td>0.85</td>
</tr>
<tr>
<td>segment</td>
<td>0.5</td>
<td>1.45</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>soybean</td>
<td>0.32</td>
<td>0.05</td>
<td>0.07</td>
<td>0.38</td>
</tr>
<tr>
<td>splice-1</td>
<td>0.21</td>
<td>30.41</td>
<td>31.11</td>
<td>35.02</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.5</td>
<td>0.85</td>
<td>&gt;</td>
<td>&gt;</td>
</tr>
<tr>
<td>yeast</td>
<td>0.49</td>
<td>5.67</td>
<td>781.63</td>
<td>&gt;</td>
</tr>
</tbody>
</table>
Experiments

- **Outside CP:**
  - DDPMine [ICDE'08]
  - LCM (FIMI's “winner”)
  - CIMCP (4-bound in Gecode CP solver)
  - corrmine (4-bound pruning implemented in a eclat-like specialised miner)
## Experiments in CP

Runtime in seconds, >900s indicated by >
memory exhausted by -

<table>
<thead>
<tr>
<th>Name</th>
<th>cormine</th>
<th>cimcp</th>
<th>ddpmine</th>
<th>lcm</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>0.02</td>
<td>0.22</td>
<td>22.46</td>
<td>7.92</td>
</tr>
<tr>
<td>australian-credit</td>
<td>0.01</td>
<td>0.30</td>
<td>3.40</td>
<td>1.22</td>
</tr>
<tr>
<td>breast-wisconsin</td>
<td>0.03</td>
<td>0.28</td>
<td>96.75</td>
<td>27.49</td>
</tr>
<tr>
<td>diabetes</td>
<td>0.36</td>
<td>2.45</td>
<td>-</td>
<td>697.12</td>
</tr>
<tr>
<td>german-credit</td>
<td>0.07</td>
<td>2.39</td>
<td>-</td>
<td>30.84</td>
</tr>
<tr>
<td>heart-cleveland</td>
<td>0.03</td>
<td>0.19</td>
<td>9.49</td>
<td>2.87</td>
</tr>
<tr>
<td>hypothyroid</td>
<td>0.02</td>
<td>0.71</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>ionosphere</td>
<td>0.24</td>
<td>1.44</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>0.02</td>
<td>0.92</td>
<td>125.60</td>
<td>25.62</td>
</tr>
<tr>
<td>letter</td>
<td>0.65</td>
<td>52.66</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>mushroom</td>
<td>0.03</td>
<td>14.11</td>
<td>0.09</td>
<td>0.03</td>
</tr>
<tr>
<td>pendigits</td>
<td>0.18</td>
<td>3.68</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>primary-tumor</td>
<td>0.01</td>
<td>0.03</td>
<td>0.26</td>
<td>0.08</td>
</tr>
<tr>
<td>segment</td>
<td>0.06</td>
<td>1.45</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>soybean</td>
<td>0.01</td>
<td>0.05</td>
<td>0.05</td>
<td>0.02</td>
</tr>
<tr>
<td>splice-1</td>
<td>0.05</td>
<td>30.41</td>
<td>1.86</td>
<td>0.02</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.07</td>
<td>0.85</td>
<td>-</td>
<td>&gt;</td>
</tr>
<tr>
<td>yeast</td>
<td>0.80</td>
<td>5.67</td>
<td>-</td>
<td>185.28</td>
</tr>
</tbody>
</table>

**avg. when found:** 0.15 6.55 28.88+ 81.54+
New bound results in far **better pruning**

CP (gecode) incurs **overhead** for very sparse datasets

Principles from CP-mining **carry back over** to traditional mining algorithms

**Fastest algorithm** in all our experiments
Parameter-free mining?

Can we do even better?

- Mine all possible itemsets for which a correlation measure exists under which it is optimal?

  = All itemsets on the convex hull in ROC space
Experiments convex hull

- No parameters
- All patterns on convex hull
- Possible!
- Reasonably small hulls
- Reasonable increase in runtime for entire hull

<table>
<thead>
<tr>
<th>Name</th>
<th>cimgp time (s)</th>
<th>cimgp convex hull time (s)</th>
<th>size of hull</th>
</tr>
</thead>
<tbody>
<tr>
<td>anneal</td>
<td>0.22</td>
<td>0.44</td>
<td>17</td>
</tr>
<tr>
<td>australian-credit</td>
<td>0.30</td>
<td>1.33</td>
<td>22</td>
</tr>
<tr>
<td>breast-wisconsin</td>
<td>0.28</td>
<td>0.83</td>
<td>20</td>
</tr>
<tr>
<td>diabetes</td>
<td>2.45</td>
<td>11.9</td>
<td>30</td>
</tr>
<tr>
<td>german-credit</td>
<td>2.39</td>
<td>3.93</td>
<td>21</td>
</tr>
<tr>
<td>heart-cleveland</td>
<td>0.19</td>
<td>0.37</td>
<td>20</td>
</tr>
<tr>
<td>hypothyroid</td>
<td>0.71</td>
<td>3.01</td>
<td>19</td>
</tr>
<tr>
<td>ionosphere</td>
<td>1.44</td>
<td>8.69</td>
<td>15</td>
</tr>
<tr>
<td>kr-vs-kp</td>
<td>0.92</td>
<td>1.75</td>
<td>17</td>
</tr>
<tr>
<td>letter</td>
<td>52.66</td>
<td>405.14</td>
<td>34</td>
</tr>
<tr>
<td>mushroom</td>
<td>14.11</td>
<td>32.45</td>
<td>10</td>
</tr>
<tr>
<td>pendigits</td>
<td>3.68</td>
<td>45.79</td>
<td>19</td>
</tr>
<tr>
<td>primary-tumor</td>
<td>0.03</td>
<td>0.07</td>
<td>16</td>
</tr>
<tr>
<td>segment</td>
<td>1.45</td>
<td>8.96</td>
<td>6</td>
</tr>
<tr>
<td>soybean</td>
<td>0.05</td>
<td>0.09</td>
<td>9</td>
</tr>
<tr>
<td>splice-1</td>
<td>30.41</td>
<td>40.13</td>
<td>10</td>
</tr>
<tr>
<td>vehicle</td>
<td>0.85</td>
<td>4.12</td>
<td>22</td>
</tr>
<tr>
<td>yeast</td>
<td>5.67</td>
<td>25.51</td>
<td>28</td>
</tr>
<tr>
<td><strong>average:</strong></td>
<td><strong>6.55</strong></td>
<td><strong>33.03</strong></td>
<td><strong>18.61</strong></td>
</tr>
</tbody>
</table>
Constraint Programming for Itemset Mining

I. Motivation, pattern mining
II. Constraint Programming basics
III. Constraint-based itemset mining using CP
IV. Correlated itemset mining using CP
V. Conclusions.
Unrelated work

Boosting / sparsity induced learning

- Every correlated itemset is a rule; a weak classifier
- LPboost [iboost: H. Saigo, T. Uno, K. Tsuda, 2007]

Statistical validation of itemsets

A new methodology for constraint-based mining

- Pattern Mining as model + search
- Using a declarative CP language
- Itemset Mining as standard depth-first search

Yet keeping the existing principles.

- Anti-monotonicity
- Similar traversal as specialized miners like eclat, dual miner, mafia, examiner, ...
Many additional advantages:

- Easily **combining** constraints
  - Demonstrated: Emerging + delta-closed + max-size + min-size

- **Studying** constraints independently
  - Demonstrated: Correlation constraint; 1-bound, 2-bound and 4-bound

- Rapid **prototyping** of new constraints
  - Demonstrated: Entire ROC convex hull
Constraint Programming for Itemset Mining

Based on open-source Gecode library for CP
- C++, very efficient, well documented
- Generic and extensible

Constraint Programming for Itemset Mining
- Also open-source and extensible
- Many constraints and documentation

CP (gecode) has overhead for sparse data
- Specialised solver with same flexibility?

Building global models *(eg. boosting)*
- Incorporate more of the learning in the mining?

In Data Mining, different pattern types and data
- graphs, trees, sequences with CP?
Bigger picture

Pattern Mining

- Efficient solvers for large binary domains
- New applications

Constraint Programming

- Technique of domains and propagation
- Flexible solvers
Efficient solvers for large binary domains
New applications

Pattern Mining

Constraint Programming

 Technique of domains
and propagation

 New general solvers

questions?