**Agglomerating Local Patterns Hierarchically with ALPHA**

Loïc Cerf  Pierre-Nicolas Mougel  Jean-François Boulicaut  
Université de Lyon, CNRS, ANR BINGO2 (MDCO 2007–2010)  
INSA-Lyon, LIRIS, UMR5205, F-69621, France

---

### From n-ary relations to agglomerated patterns

<table>
<thead>
<tr>
<th>0111</th>
<th>0110</th>
<th>1100</th>
<th>1110</th>
<th>1000</th>
</tr>
</thead>
</table>

### A three-step method

1. Complete collections of (possibly error-tolerant) closed patterns are extracted from arbitrary n-ary relations ($n \geq 2$);  
2. They are hierarchically agglomerated;  
3. The relevant agglomerated patterns are selected.

---

### Step 1: Extraction of patterns from arbitrary n-ary relations ($n \geq 2$)

**FENSTER:**
- generalizes error-tolerant closed set mining;  
- handles a very broad class of constraints;  
- computes complete collections of error-tolerant patterns, called closed ET-$n$-sets, from arbitrary n-ary relations (bounded number of errors per element in the pattern);  
- cannot tolerate enough errors while remaining tractable on both large and noisy relations.

**ALPHA** does not require the initial collection to present particular properties (e.g., closedness). However, a complete collection looks trustier (lossy heuristics delayed as far as possible in the knowledge discovery process).

---

### Step 2: Hierarchical agglomeration of the n-sets

**Definition (Metric)**

Given a weight function $w : \mathcal{P}_n \times \mathcal{P}_n \rightarrow \mathbb{R}^+$ (greater weights mean less tolerance to noise) and an $n$-set $X = (X^1, \ldots, X^n)$, $d(X)$ denotes its intrinsic distance:

$$d(X) = \max_{i=1}^{n} \left( \max_{x \in X^i} \left( \frac{w(x) |K \setminus R|}{|K|} \right) \right),$$

where $K = X^1 \times \cdots \times X^{i-1} \times \{x\} \times X^{i+1} \times \cdots \times X^n$.

The distance between two $n$-sets $X$ and $Y$ is $d(X \cup Y)$. Thus, unlike most pattern clustering approaches, all the information inside the minimal envelope of the two agglomerated patterns is taken into account.

**Intuition**

Restricted to the two $3 \times 3$ closed 2-sets they contain, the two binary settings on the right are identical. However, the patterns of the rightmost setting obviously are better candidates for an agglomeration. The (unweighted) metric of ALPHA reflects that: the two closed 2-sets are distant by $\frac{1}{2}$ in this setting, whereas they are distant by $\frac{1}{4}$ in the other setting.

---

### Step 3: Selection of the relevant agglomerated patterns

**Definition (Relevancy)**

Given an $n$-set $X$ and its parent $X \cup Y$ after hierarchical agglomeration, $r(X)$ denotes the relevancy of $X$:

$$r(X) = d(X \cup Y) - \max(d(X), d(Y)) - d(X)$$

**Intuition**

The relevancy measure follows the definition of a local pattern by David Hand:  
A local pattern is a data vector serving to describe an anomalously high local density of data points.

---

### Perspectives

The list of agglomerated patterns is read by decreasing relevancy order. ALPHA tests whether the previously browsed patterns cover (i.e., are ancestors of) every $n$-set extracted at Step 1. Once this assertion is true, the remaining patterns are erased.

**Intuition**

The completeness of Step 1 is somehow preserved: every extracted $n$-set is part of at least one output pattern.