Complete Discovery of High-Quality Patterns in Large Numerical Tensors

UFMG - ICEx
DEPARTAMENTO DE CIÊNCIA DA COMPUTAÇÃO

Loïc Cerf Wagner Meira Jr.

Numerical tensor $\mathcal{T} \in \mathbb{R}^{\prod_{i=1}^{n} \mathcal{D}_i}$

Numerical tensors associate n-tuples with numerical values:

$$\begin{array}{lll} \{ & (\texttt{week 1}, \texttt{Botafogo}, \texttt{3toques}) & \mapsto \text{6 retweets} \\ & (\texttt{week 1}, \texttt{Botafogo}, \texttt{4evertton}) & \mapsto \text{5 retweets} \\ & \vdots & \vdots \\ & (\texttt{week 45}, \texttt{Corinthians}, \texttt{espnagora}) & \mapsto \text{185 retweets} \\ & \vdots & \vdots \\ & (\texttt{week 49}, \texttt{Goiás}, \texttt{milton\_neves}) & \mapsto \text{100 retweets} \\ & \vdots & \vdots \\ & (\texttt{week 49}, \texttt{Vasco}, \texttt{zigugo28}) & \mapsto \text{1 retweet} \quad \} \end{array}$$

Pattern $X \in \prod_{i=1}^{n} 2^{\mathcal{D}_i}$

A pattern binds subsets of each of the n dimensions. With an appropriate semantics, it supports the discovery of correlations between elements of all dimensions:

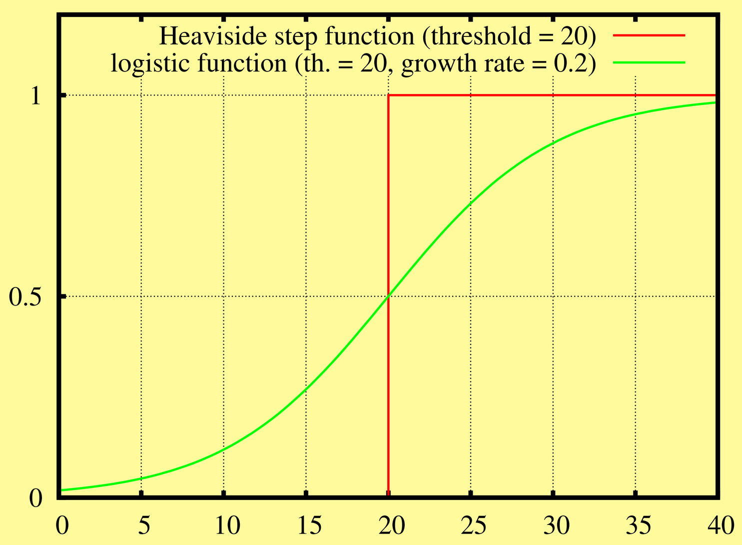
$$(\{\texttt{week 45}, \texttt{week 47}, \texttt{week 48}, \texttt{week 49}\}, \{\texttt{Corinthians}, \texttt{Fluminense}, \texttt{Goiás}\}, \{\texttt{espnagora}, \texttt{ge\_timao}, \texttt{milton\_neves}\})$$

The pattern is said to *cover* the Cartesian product of its n subsets. E. g., the pattern above covers this 3-tuple:

$$(\texttt{week 45}, \texttt{Corinthians}, \texttt{ge\_timao})$$

Fuzzy relation $\mathcal{R} \in [0,1]^{\prod_{i=1}^{n} \mathcal{D}_i}$

To define patterns covering n-tuples associated with large numbers (for instance), the numerical tensor is turned into a fuzzy relation. E. g., a logistic function is applied:



The extremal Heaviside step function is commonly used but implies a loss of information w.r.t. the numerical tensor.

Absolute and per-element tolerance to noise

The *noise* a pattern tolerates is a distance between the membership degrees of its cover in the fuzzy relation and '1' values. In our opinion, noise tolerance must be:

absolute so that 1) the sole closed patterns form a lossless representation of all patterns and 2) the tolerance is anti-monotone, hence allows a scalable search of the valid patterns in the pattern space.

per-element so that, contrary to a per-pattern tolerance, every element really belongs to the pattern. E. g., in the relation below, $(\{\alpha,\beta,\gamma\},\{a,b,c,d\})$ must be invalid:

	a	b	c	d
α	0.4	1	1	0
β	1	1	1	0
γ	1	1	1	0

Definition of a valid pattern

Given n noise tolerance thresholds $\epsilon = (\epsilon_i)_{i=1..n} \in \mathbb{R}_+^n$, the pattern $X = (X_i)_{i=1..n} \in \prod_{i=1}^{n} 2^{\mathcal{D}_i}$ is valid if and only if:

$\mathcal{C}_{\epsilon\text{-connected}}$ $\forall i = 1..n, \forall x \in X_i, \sum_{t \in X_1 \times \cdots \times \{x\} \times \cdots \times X_n} 1 - \mathcal{R}_t \leq \epsilon_i$

$\mathcal{C}_{\epsilon\text{-closed}}$ $\forall i = 1..n, \forall s \in \mathcal{D}_i \setminus X_i,$

$$\begin{cases} \sum_{t \in X_1 \times \cdots \times \{s\} \times \cdots \times X_n} 1 - \mathcal{R}_t > \epsilon_i \\ \text{or} \\ \exists x \in \cup_{j \neq i} X_j \mid \sum_{t \in X_1 \times \cdots \times X_i \cup \{s\} \times \cdots \times \{x\} \times \cdots \times X_n} 1 - \mathcal{R}_t > \epsilon_j \end{cases}$$

Any other relevant property is enforced via an additional user-defined constraint $\mathcal{C}$. To have it prune the pattern space, each *occurrence* of the pattern in its expression must behave monotonically or anti-monotonically. E. g., this constraint forces the patterns, with a numerical dimension $X_i \subset \mathbb{R}$, to have those elements τ-close from each other:

$$\forall t \in [\min(X_i), \max(X_i)], \exists t' \in X_i \text{ s.t. } |t' - t| \leq \tau$$

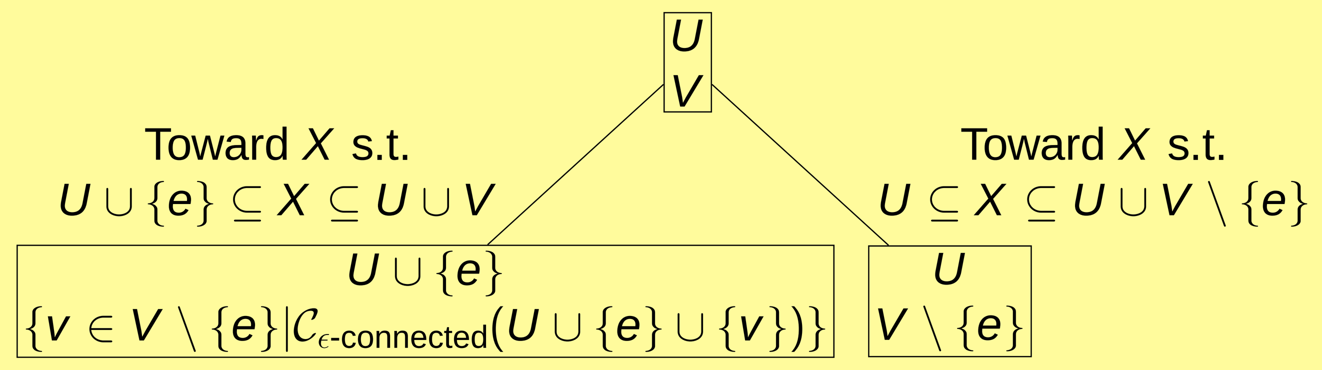
Traversal of the pattern space

A binary tree structures the pattern space. At every node, root of a sub-tree:

$U \subseteq \cup_{i=1}^{n} \mathcal{D}_i$ contains the elements that *must* be present in every pattern at the leaves of the sub-tree;

$V \subseteq \cup_{i=1}^{n} \mathcal{D}_i$ contains the elements that *can* be present in some patterns at the leaves of the sub-tree.

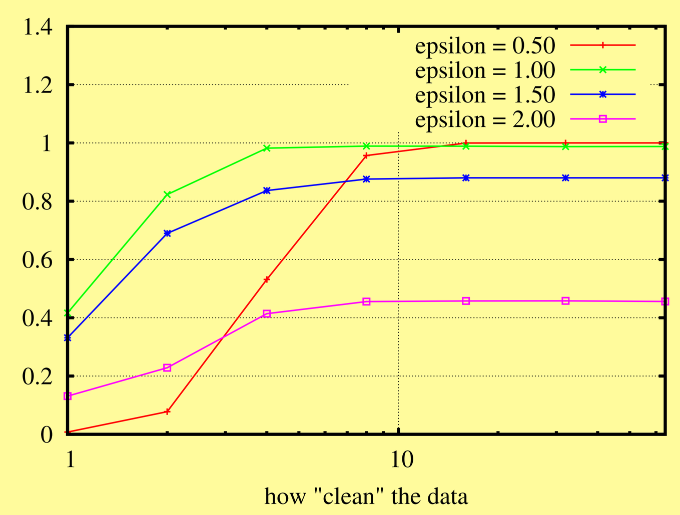
At the root, $(U, V) = (\emptyset, \cup_{i=1}^{n} \mathcal{D}_i)$. A reduced tree, which does not miss any valid pattern, is traversed depth-first. If $V \neq \emptyset$, $e \in V$ is freely chosen and these rules applied:



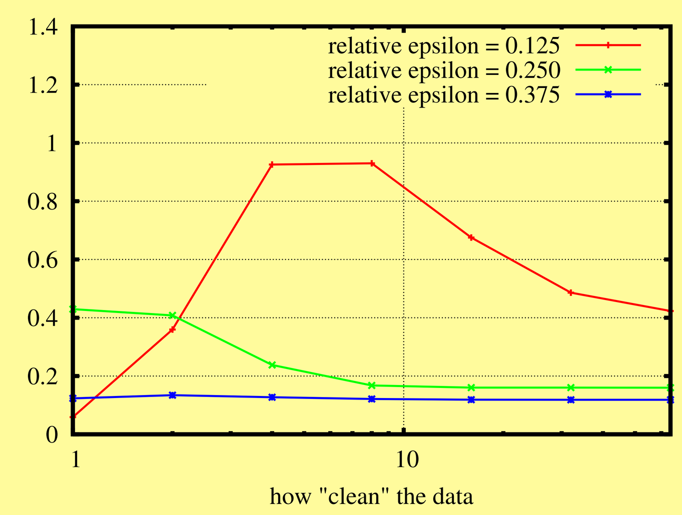
$\mathcal{C}_{\epsilon\text{-closed}}$ and $\mathcal{C}$ are enforced at every node (pruning).

Quality of the pattern collection ($\mathcal{R}$ is $16 \times 16 \times 16$)

Per-element tolerance

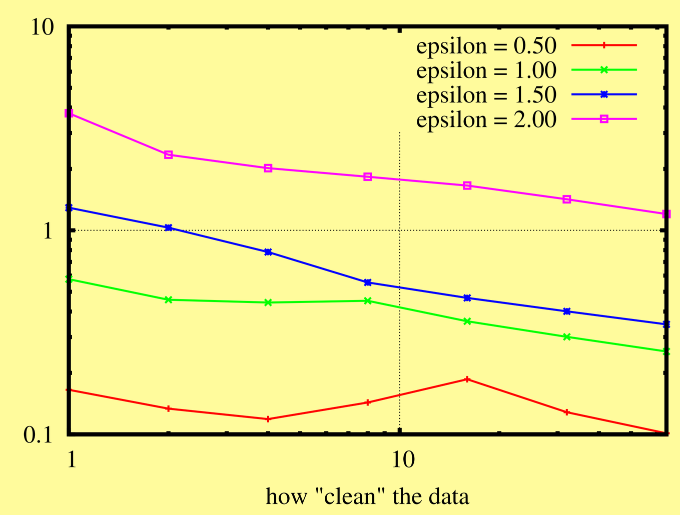


Per-pattern tolerance (DCE)



Running time in seconds ($\mathcal{R}$ is $16 \times 16 \times 16$)

Absolute tolerance



Relative tolerance (DCE)