Mining Skypatterns in Fuzzy Tensors

dupe hacks in the sky

Nicolas Nadisic **Aurélien Coussat** **Loïc Cerf**
Universidade Federal de Minas Gerais (lcerf@dcc.ufmg.br)

Objectives

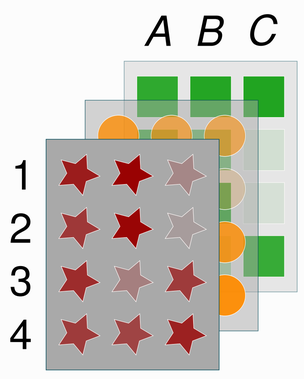
- Generalize skypattern mining from binary matrices to fuzzy tensors.
- Efficient/fast algorithm.
- Manage a wide class of measures.

Fuzzy tensors

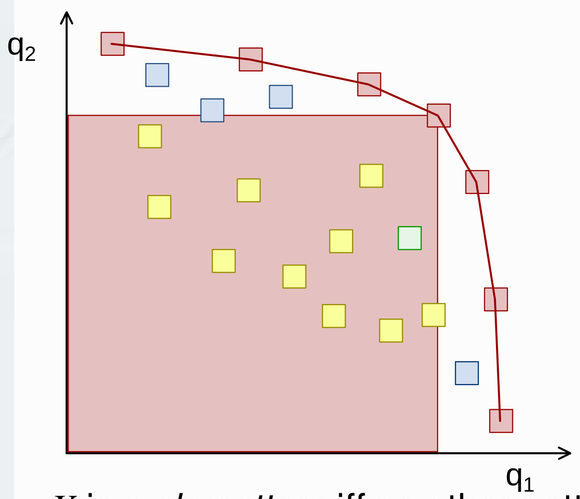
	A	B	C	...
1	0.1	0	1	...
2	1	0.9	0.2	...
3	0.8	0.8	0.1	...
4	0.8	1	0.3	...
⋮	⋮	⋮	⋮	

Fuzzy matrix

- Noise-Tolerance per dimension.
- Ex $\epsilon_{\text{row}} = 0.4$, $\epsilon_{\text{col}} = 0.3$.
- Every ϵ_i upper-bounds the absolute amount of noise in every "slice" of the pattern that relates to the elements in its i^{th} dimension.



Skypatterns



- Pareto-optimal patterns.
- Optimize simultaneously various measures.
- No need for user-defined thresholds.

Existing algorithms:

- `aetheris`
- `CP+Sky`
- `dominance programming`

X is a *skypattern* iff no other pattern scores at least as well w.r.t. every measure and strictly better w.r.t. one of them.

Definitions

Given the dimensions $D_1, \ldots, D_n$ of the fuzzy tensor:

Definition (Rewriting)

m' is a *rewriting* of a measure m if and only if it is a function, whose domain is $\left(\prod_{i=1}^{n} 2^{D_i}\right)^2$, whose codomain is $\mathbb{R} \cup \{+\infty\}$, and that is such that $\forall X \in \prod_{i=1}^{n} 2^{D_i}$, $m'(X, X) = m(X)$.

Definition (Piecewise (anti-)monotonicity)

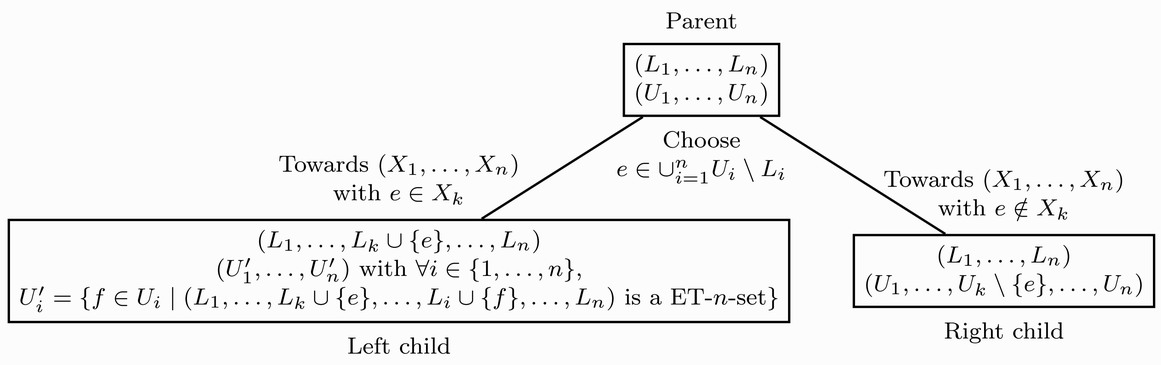
A measure m is said *piecewise (anti-)monotone* if and only if there exists a rewriting m' of m such that:

$$\forall U \in \prod_{i=1}^{n} 2^{D_i}, \forall X \in \prod_{i=1}^{n} 2^{U_i}, \forall L \in \prod_{i=1}^{n} 2^{X_i}, m(X) \leq m'(L, U) \quad .$$

Example (Growth-rate to a subset $C \subseteq D_1$)

$$(X_1, X_2) \mapsto \frac{|D_1 \setminus C| \times |X_1 \cap C|}{|C| \times |X_1 \setminus C|} \Leftrightarrow (L_1, L_2, U_1, U_2) \mapsto \frac{|D_1 \setminus C| \times |U_1 \cap C|}{|C| \times |L_1 \setminus C|}$$

`multidupehack` pattern space traversal

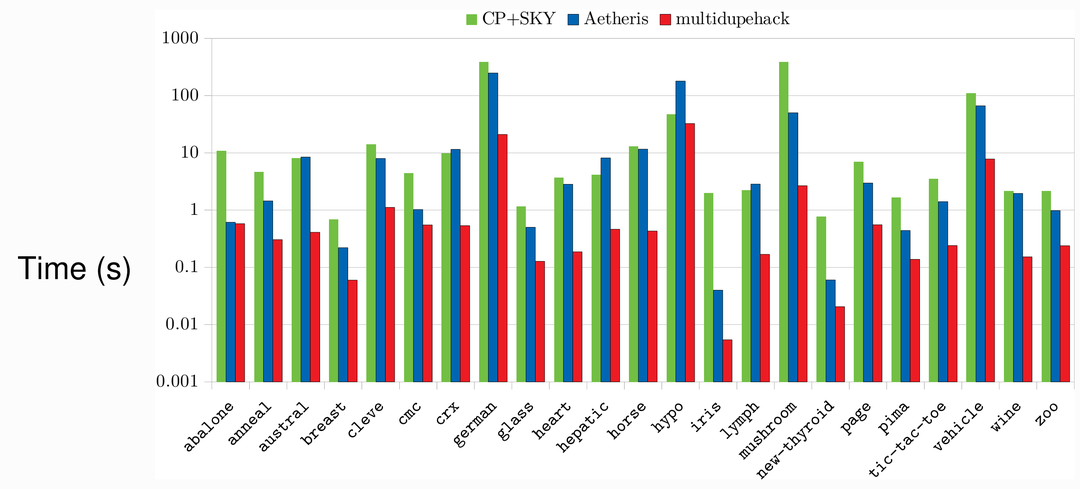


`multidupehack` for skypattern mining

Data: $T, \epsilon_1, \ldots, \epsilon_n, M'$
Result: the skypatterns in T
$\mathcal{S} \leftarrow \emptyset$ `mine`$(\emptyset, \ldots, \emptyset, D_1, \ldots, D_n)$
return $\mathcal{S}$

Function *mine*(L, U)**:**
  if $\forall P \in \mathcal{S}$,
  $\exists m' \in M' \mid m'(P, P) < m'(L, U) \vee \forall m' \in M', m'(P, P) \leq m'(L, U)$ **then**
    if $L = U$ **then**
      $\mathcal{S} \leftarrow \{P \in \mathcal{S} \mid L \not\succ_M P\} \cup \{L\}$
    else
      choose $e \in \cup_{i=1}^{n} U_i \setminus L_i$
      `/* Let` k `the index of the dimension of` e `*/`
      `mine`$(L_1, \ldots, L_k \cup \{e\}, \ldots, L_n, U'_1, \ldots, U'_n)$
      `/* where` U'_i `is defined as above figure */`
      `mine`$(L_1, \ldots, L_n, U_1, \ldots, U_k \setminus \{e\}, \ldots, U_n)$
    end
  end

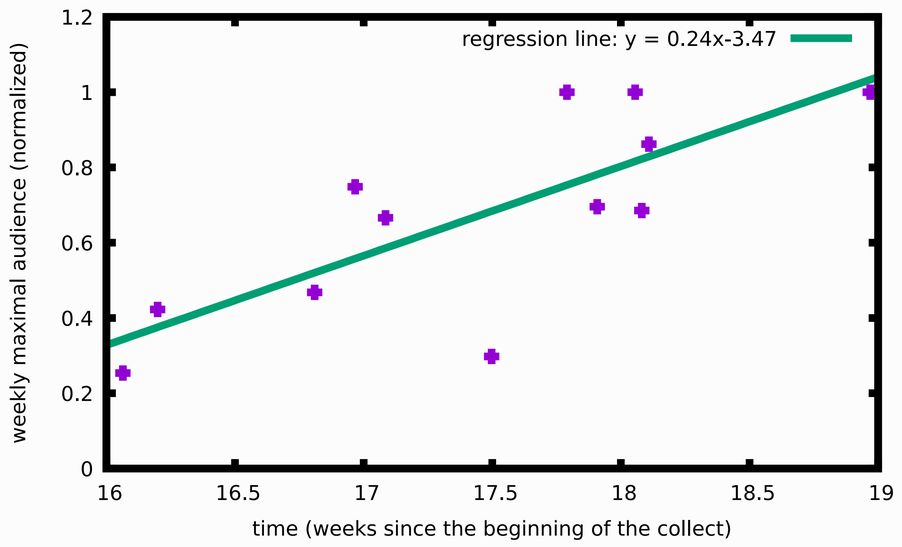
Comparison on UCI datasets



Extracting skypatterns maximizing freq, area, and growth-rate.

Twitch: real-life 3-way fuzzy tensor

- Connection times of spectators of Starcraft games streaming, from October 2013 to February 2014.
- 1,198,292 spectators, 92 channels, 19 weeks.
- To each pair (channel,week) is associated a 2D point (ordinate is the maximal number of simultaneous viewers; abscissa is the time this maximum was reached).
- Extract skypatterns maximizing number of channels, number of weeks, and the slope of the regression line of the 2D points.



Results

- 114 skypatterns extracted in 7 minutes 34 seconds.
- High relevancy.
- Ex:

$(\{$`balosaar, darrenlorduk, japi1, lumberdavid, yoctz22, 91kitsune91`$\}$,
$\{$`egjd, egstephano, esltv_sc2, sc2proleague`$\}$,
$\{$`week17, week18, week19`$\})$