This thesis is funded by the ANR and the Region Normandie as part of the HAISCoDe project.

Efficiently Sampling Interval Patterns from Numerical Databases

Djawad Bekkoucha[1], Lamine Diop[2], Abdelkader Ouali[1] , Bruno Crémilleux[1] and Patrice Boizumault[1]

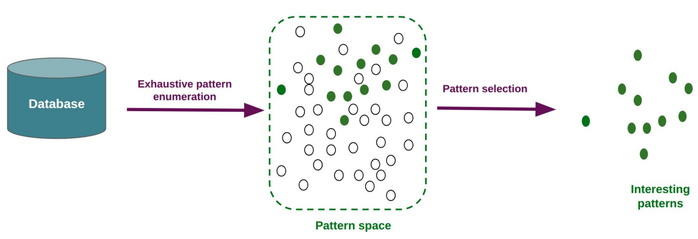
[1] University of Caen Normandy, ENSICAEN, CNRS, Normandy Univ GREYC UMR6072, F-14000 Caen, France
[2] EPITA Research Laboratory (LRE), Le Kremlin-Bicêtre, Paris, FR-94276, France

djawad.bekkoucha@unicaen.fr

GREYC
Electronics and Computer Science Laboratory

1. Approaches for Pattern Discovery

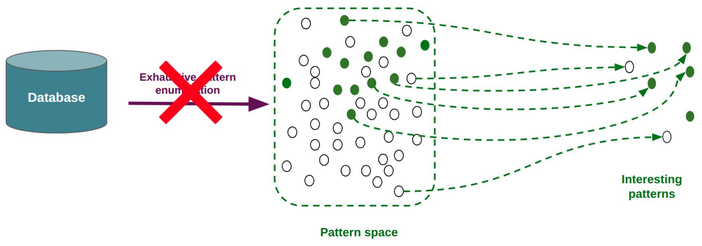
Exhaustive pattern mining



Drawbacks:

- **High computational cost**
- **Large number of patterns**
- **Pattern redundancy**

Pattern sampling



Advantages:

- **Low computational cost**
- **Control on the resulting number of patterns**
- **Controlled sampling distribution**

2. How to mine patterns from numerical data?

- Numerical databases are widely used in a large number of fields
- Use the interval pattern language to represent informations from numerical data:

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

	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

Example

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

3. State of the Art

Stochastic methods:

- **Graphs:** [El Hasan et al.VLDB 2009], [Bendimerad et al., ICDM 2016]
- **Formal concepts:** [Boley et al., DSM 2010]
- **Tiles:** [Bendimerad et al., IDA 2020]

Déclarative method:

- **Itemsets:** [Dzyuba et al., DAMI 2017]

No sampling approach for interval patterns

Multisteps methods:

- **Itemsets:** [Boley et al., KDD 2011]
- **Sequences:** [Diop et al., ICDM 2018]
- **Numerical data:** [Giacometti et al., SDM 2018]

4. Frequency-Based Interval Pattern Sampling

Pre-processing

- **Goal:** compute the sum of frequencies of all interval patterns in search space **without enumeration**
- Equivalent to counting the number of patterns covering each object $g \in \mathcal{G}$:

$$NIP(g) = \prod_{m \in \mathcal{M}} |\mathcal{U}(v_{g,m})| + |\mathcal{A}(v_{g,m})| + |\mathcal{U}(v_{g,m})| \cdot |\mathcal{A}(v_{g,m})| + 1$$

such that $\begin{cases} \mathcal{U}(v_{g,m}) = \{v \in \mathcal{N}_m \mid v < v_{g,m}\} \\ \mathcal{A}(v_{g,m}) = \{v \in \mathcal{N}_m \mid v > v_{g,m}\} \end{cases}$

Example:
Let us consider attribute m1 and object g1. The value 155 can be:

- A lower bound : $[155, 167],\ [155, 176],\ [155, 190]$
- An upper bound : $[153, 155]$
- Strictly included : $[153, 167],\ [153, 176],\ [153, 190]$
- Both lower and upper bounds : $[155, 155]$

	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

The number of intervals for m1 including 155 is equal to 8

Step 1

- Draw an object proportionally to its NIP value

	Height m_1	Weight m_2	Age m_3	$NIP(g)$
g_1	155	74	80	240
g_2	176	99	74	288
g_3	167	73	28	180
g_4	153	76	52	240
g_5	190	99	76	260

- The object g_2 is more likely to be sampled

Step 2

- Draw an interval pattern uniformly

	Height m_1	Weight m_2	Age m_3	$NIP(g)$
g_1	155	74	80	240
g_2	176	99	74	288
g_3	167	73	28	180
g_4	153	76	52	240
g_5	190	99	76	260

<[155, 190], [74, 99], [74, 76]>

Theoretical Results

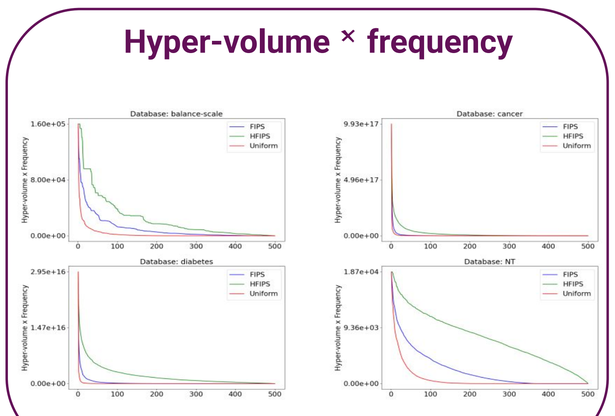
- Extending FIPS with hyper-volume using the IPH function:

$$IPH(g) = \prod_{m \in \mathcal{M}} \left(\sum_{v \in J(v_{g,m})} x \cdot |I(v_{g,m})| \right) - \left(\sum_{v \in I(v_{g,m})} x \cdot |J(v_{g,m})| \right) \quad \text{such that} \begin{cases} I(v_{g,m}) = \{v \in \mathcal{N}_m \mid v \leq v_{g,m}\} \\ J(v_{g,m}) = \{v \in \mathcal{N}_m \mid v \geq v_{g,m}\} \end{cases}$$

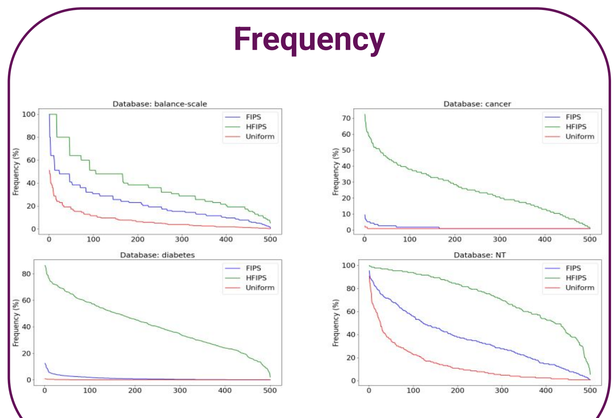
- Theoretical proofs for the sampling distributions

5. Experimental Results

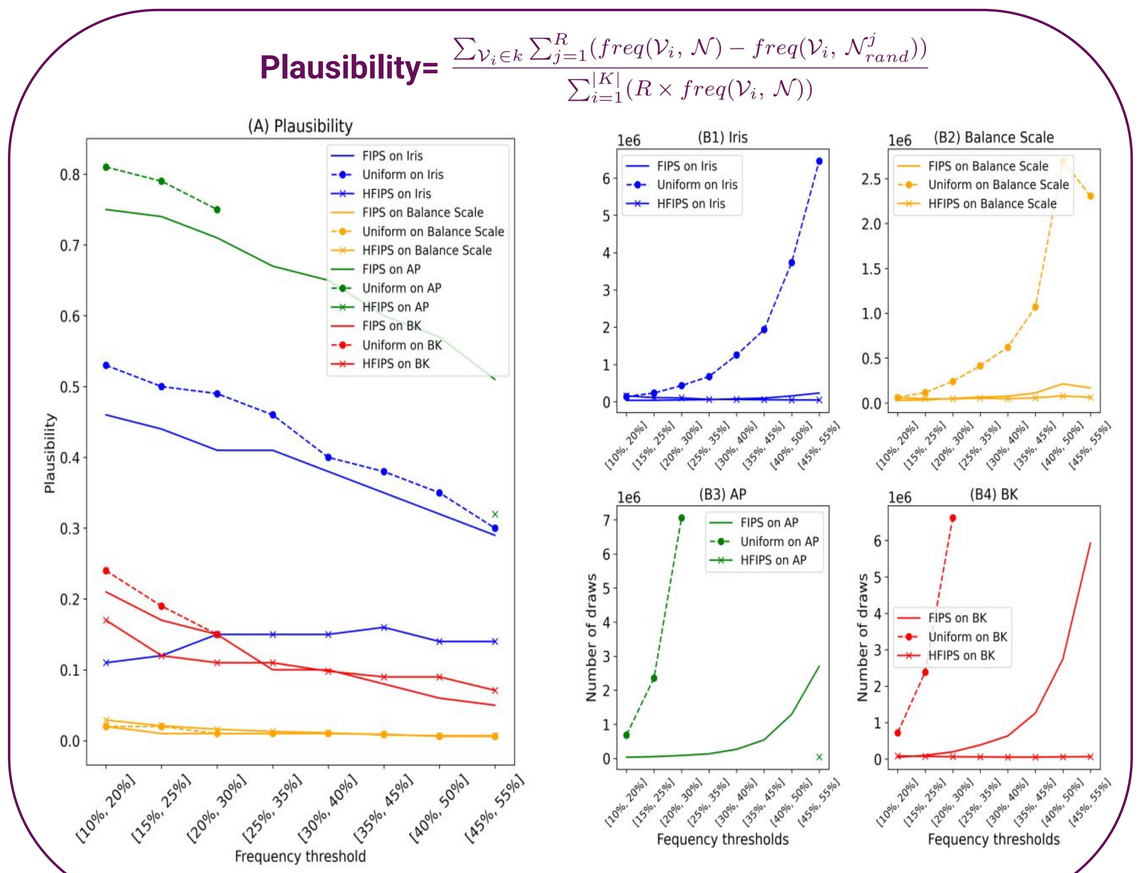
Hyper-volume × frequency



Frequency

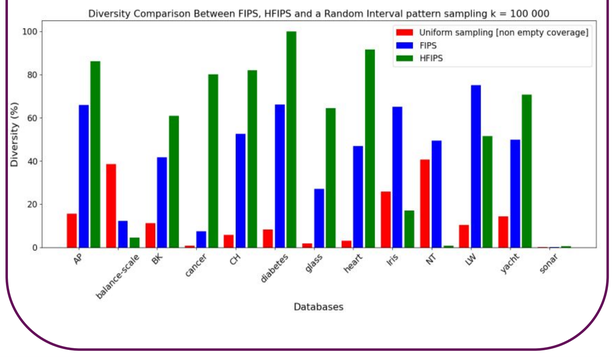


Plausibility= $\dfrac{\sum_{\mathcal{V}_i \in k} \sum_{j=1}^{R} (freq(\mathcal{V}_i, \mathcal{N}) - freq(\mathcal{V}_i, \mathcal{N}^j_{rand}))}{\sum_{i=1}^{|K|} (R \times freq(\mathcal{V}_i, \mathcal{N}))}$

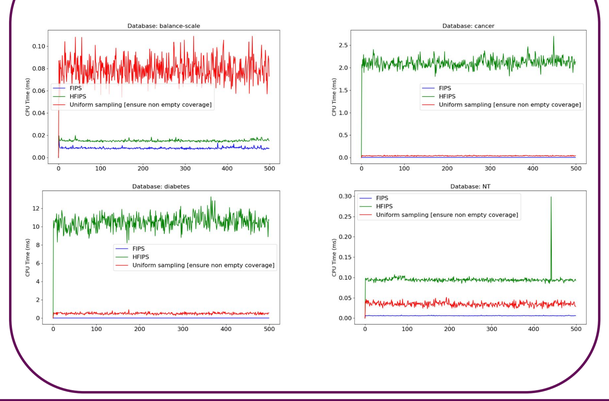


Diversity

$$\textbf{Diversity=} \frac{|\{cover(\mathcal{V}_1, \mathcal{N}), \dots, cover(\mathcal{V}_{|K|}, \mathcal{N})\}|}{|K|}$$



CPU Time



6. Future Directions

- **Extending our approach to other interestingness measures like density**
- **Sampling patterns under constraints**
- **Incorporating our approach in interactive mining approaches**
- **Exploring the use of sampling in machine learning contexts**