KU LEUVEN

Learning what matters

Sampling interesting patterns

Vladimir Dzyuba, Matthijs van Leeuwen

KU Leuven, Belgium

PAKDD – May 26, 2017

Exploratory data analysis

Tell me (analyst/user) something interesting about the data

	Louise	Emma	Marie	Olivia	Nora	*Points scored?*
1	*Plays*	P	P	*Rests*	P	+
2	P	P		P	P	−
3	P	P	P			+
4	P			P		−
5		P		P	P	−

Exploratory data analysis with pattern mining

Discovering descriptions of coherent *interesting* data regions

Pattern p	$\lvert p \rvert$	$\lvert p^+ \rvert$	$\lvert p^- \rvert$	
$\{$Louise, Emma$\}$	3	2	1	Frequent
$\{$Marie$\}$	2	2	0	Discriminative (only $+$)
$\{$Louise, Emma, Marie$\}$	2	2	0	Discriminative & long

Exploratory data analysis with pattern mining

Problem statement

Pattern p	$\lvert p \rvert$	$\lvert p^+ \rvert$	$\lvert p^- \rvert$	
$\{$Louise, Emma$\}$	3	2	1	Frequent
$\{$Marie$\}$	2	2	0	Discriminative (only $+$)
$\{$Louise, Emma, Marie$\}$	2	2	0	Discriminative & long

Given dataset $\mathcal{D}$, constraints $\mathcal{C}$, and/or quality measure φ
Mine top patterns according to φ

Exploratory data analysis with pattern mining

Problem statement

Pattern p	$\|p\|$	$\|p^+\|$	$\|p^-\|$	
$\{$Louise, Emma$\}$	3	2	1	Frequent
$\{$Marie$\}$	2	2	0	Discriminative (only $+$)
$\{$Louise, Emma, Marie$\}$	2	2	0	Discriminative & long

Given dataset $\mathcal{D}$, constraints $\mathcal{C}$, and/or quality measure φ
Mine top patterns according to φ

Exploratory data analysis with pattern mining

Problem statement

Pattern p	$\lvert p \rvert$	$\lvert p^+ \rvert$	$\lvert p^- \rvert$	
$\{$Louise, Emma$\}$	3	2	1	Frequent
$\{$Marie$\}$	2	2	0	Discriminative (only $+$)
$\{$Louise, Emma, Marie$\}$	2	2	0	Discriminative & long

Given dataset $\mathcal{D}$, constraints $\mathcal{C}$, and/or quality measure φ
Mine top patterns according to φ

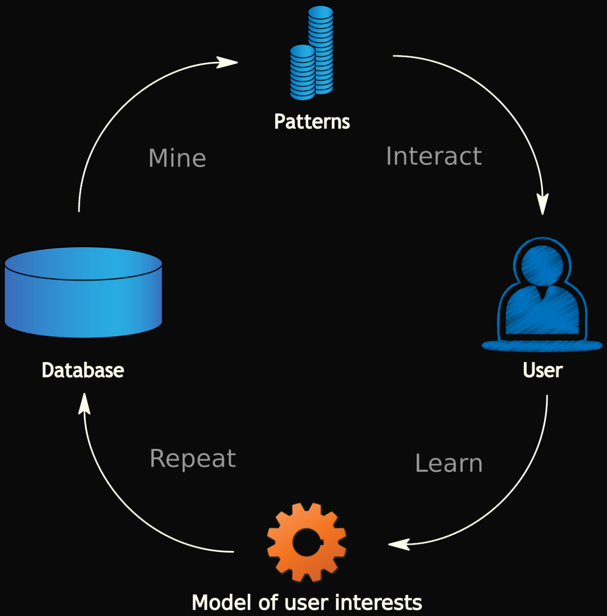
Which ones are interesting to the given user?

Subjectivity in exploratory data analysis

Pattern p	$\lvert p \rvert$	$\lvert p^+ \rvert$	$\lvert p^- \rvert$	
$\{$Louise, Emma$\}$	3	2	1	Opponent
$\{$Marie$\}$	2	2	0	Coach
$\{$Louise, Emma, Marie$\}$	2	2	0	Journalist

1. What patterns are interesting depends on the user
2. Non-experts can't tune constraints $\mathcal{C}$ or quality measure φ

Direct user involvement is essential



Interact: Ordered feedback

The user ranks patterns by their *subjective* interestingness

$\{$Louise, Emma$\}$, $\{$Louise, Emma, Marie$\}$, $\{$Marie$\}$

$$\Downarrow$$

1. $\{$Louise, Emma, Marie$\}$
2. $\{$Marie$\}$
3. $\{$Louise, Emma$\}$

Dzyuba, van Leeuwen, Nijssen, De Raedt. (2014) *Interactive learning of pattern rankings*, IJAIT

Learn: Pattern ranking/scoring function

Instance of *preference learning*

$$\varphi_{\text{user}}\left(p\right) = A + \frac{1 - A}{1 + e^{-\vec{w}\cdot\vec{p}}}$$

$\vec{p}$ Pattern features (e.g., items, frequency, length…)

$\vec{w}$ Which features make a pattern interesting to the user

Learned from ordered feedback with *stochastic coordinate descent*

A Technical parameter, see the paper

Mine: Requirements for pattern sets shown to the user

Interesting	according to the current (approximate) φ_{user}
Compact	so that feedback is easy to provide
Diverse	to ensure *exploration* necessary for learning
Quick to obtain	e.g., in an *anytime* manner

Mine: Requirements for pattern sets shown to the user

Interesting according to the current (approximate) φ_{user}

Compact so that feedback is easy to provide

Diverse to ensure *exploration* necessary for learning

Quick to obtain e.g., in an *anytime* manner

Given dataset $\mathcal{D}$, constraints $\mathcal{C}$, and quality measure φ
Sample patterns proportional to φ_{user}

Mine: Sample with *Flexics*

Efficient black-box pattern sampler with strong performance guarantees

To sample one pattern (at least):

1. Generate an *implicit* random partitioning of all patterns
2. Enumerate all patterns in a random partition
3. Generate a perfect sample from this partition

Dzyuba, van Leeuwen, De Raedt. (2017) *Flexible constrained sampling with guarantees for pattern mining*, DMKD

Illustration: all patterns

PCA representation: similar patterns are close to each other

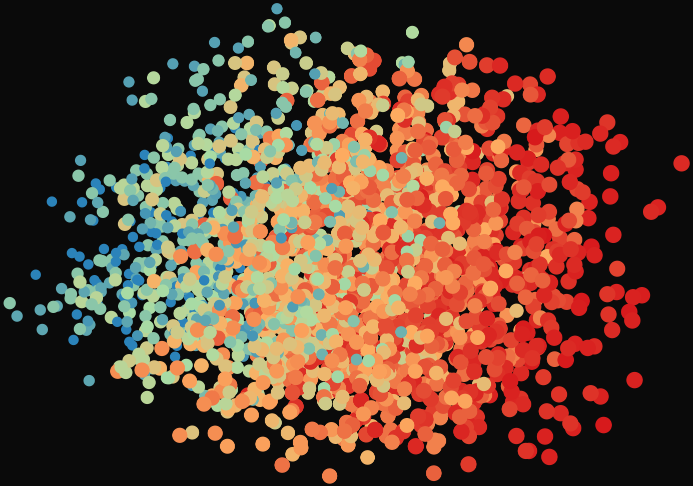
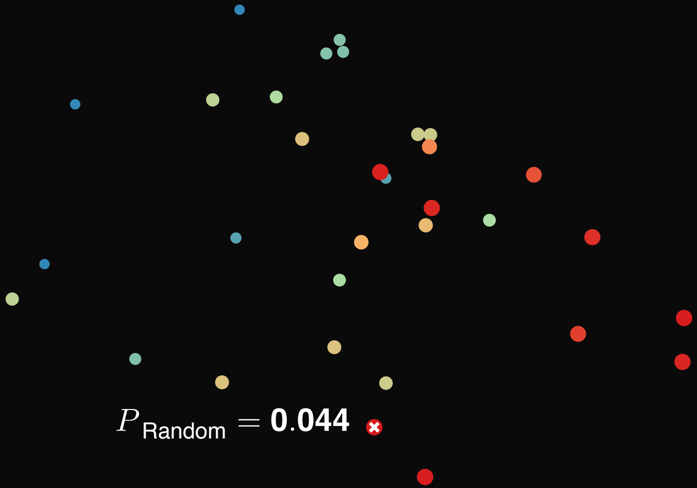


Illustration: random partition (“cell”)

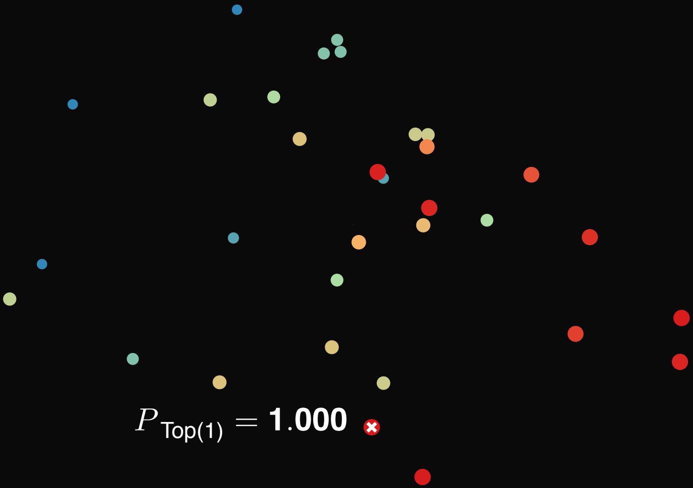
Patterns are different from each other ⇒ Good *exploration*



$P_{\text{Random}} = \mathbf{0.044}$ ✖

Illustration: random partition

Take top patterns from the partition ⇒ Emphasize *exploitation*



$P_{\text{Top(1)}} = \mathbf{1.000}$ ✗

LetSIP: End-to-end interactive pattern sampler

More contributions & details in the paper

Mine	Sample with *Flexics*
	Modified “cell” sampling
Interact	Ordered feedback
Learn	Learning φ_{user} with SCD
Repeat	Sample with updated φ_{user}

Experimental evaluation

Emulating the user by a quality measure

Frequency, surprisingness, discriminativity with χ^2

		φ_{obj}
1	$\{$Louise, Emma, Marie$\}$	0.5
2	$\{$Marie$\}$	0.3
3	$\{$Louise, Emma$\}$	0.1

Measuring performance by regret

Values of φ_{obj} are not known to the learner, only used to measure its performance

		φ_{obj}
1	$\{\text{Louise}, \text{Emma}, \text{Marie}\}$	0.5
2	$\{\text{Marie}\}$	0.3
3	$\{\text{Louise}, \text{Emma}\}$	0.1

Regret w.r.t. max.$\varphi = 1 - 0.5 \qquad = 0.5$

avg.$\varphi = 1 - (0.5 + 0.3 + 0.1)/3 \qquad = 0.3$

Experimental setup

- Learn to sample from φ_{obj} only from small orders
- 10 datasets; choose *min.frequency* so that there are 140 000+ frequent patterns
- 30 learning iterations, 5 patterns per iteration
  Regret $\in [0,\ 30]$, lower is better

Results: Exploration/exploitation

Modified cell sampling improves the performance

		Regret w.r.t.		
		Avg.qual	Max.qual	Diversity
Cell	*Random*	10.6 ± 0.7	1.9 ± 0.6	**12.2 ± 0.6**
sampling	*Top(1)*	**5.1 ± 1.1**	**0.8 ± 0.5**	13.7 ± 1.0

Results: Outperforms state-of-the-art

		Regret w.r.t.		
		Avg.qual	Max.qual	Diversity
Algorithm	LetSIP	$\mathbf{3.1 \pm 0.8}$	$\mathbf{0.3 \pm 0.2}$	$\mathbf{13.1 \pm 0.8}$
	APLe	3.2 ± 2.6	2.6 ± 2.4	–
	IPM	12.9 ± 2.4	5.1 ± 2.3	16.0 ± 1.9

Take-away messages

See the paper for further technical details and experimental results

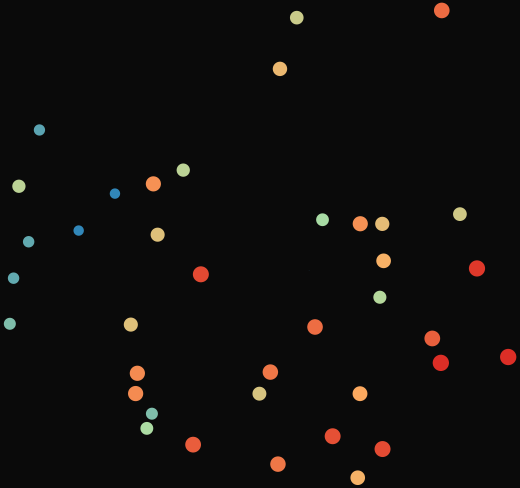
Mine, Interact, Learn, Repeat

- Pattern sampling with *Flexics* delivers compact diverse sets of high-quality patterns in an anytime fashion…
- …which helps balance exploration and exploitation in interactive mining with *LetSIP*

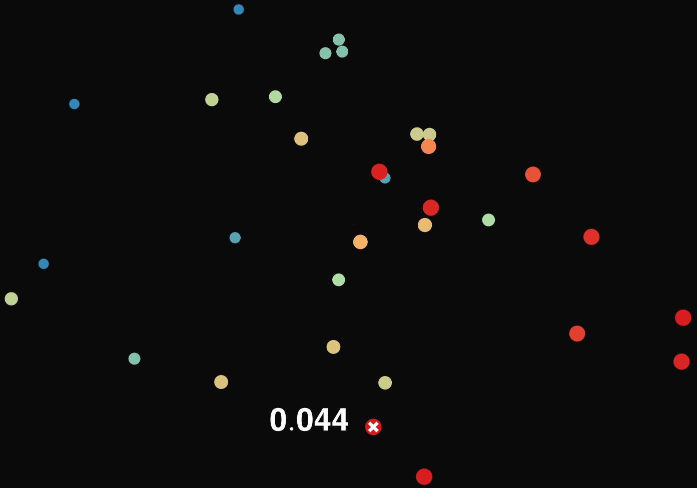
Learning what matters — Sampling interesting patterns

Thank you for your attention!

May I answer any questions?



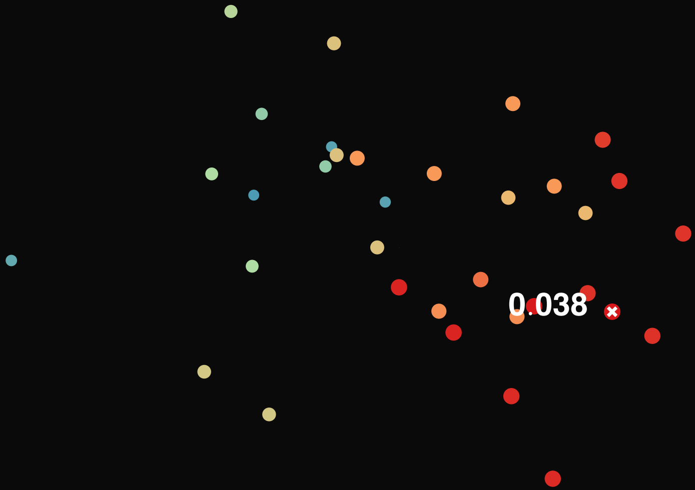
0.034 ✖



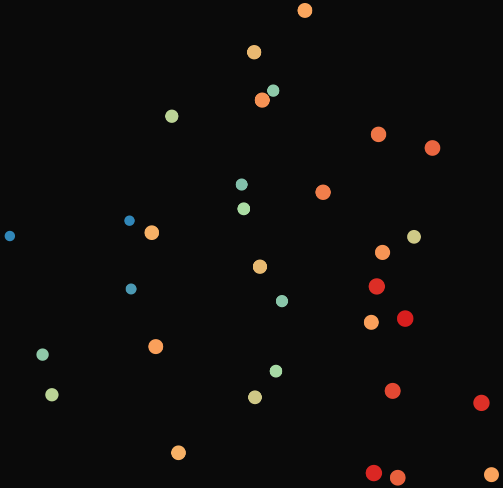
0.044

0.039

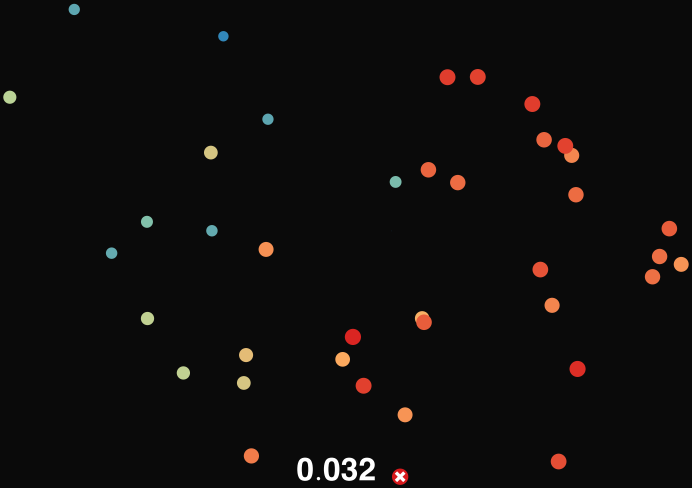
0.034

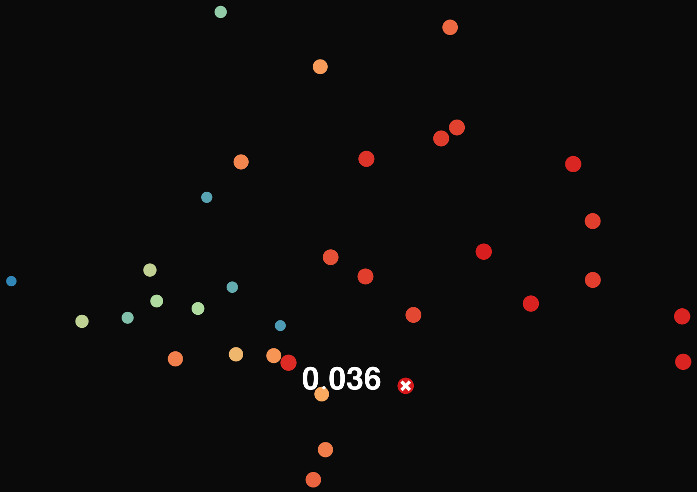

0.038

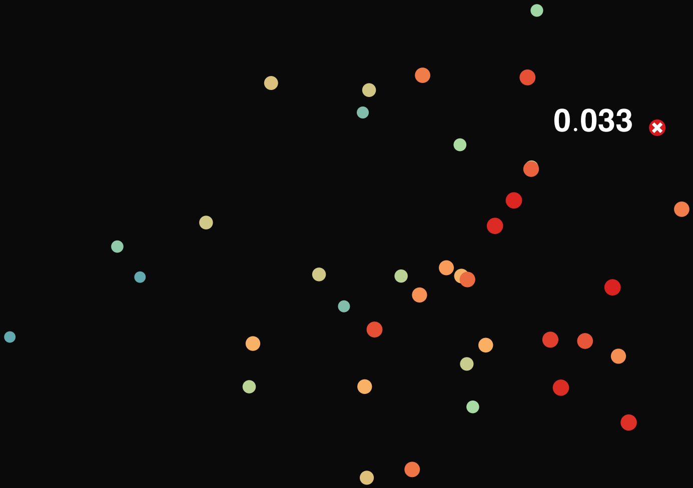
0.036



0.041


0.032


0.036


0.033