Statistical Emerging Pattern Mining with Multiple Testing Correction

Junpei Komiyama[1], Masakazu Ishihata[2],
Hiroki Arimura[2], Takashi Nishibayashi[3],
Shin-Ichi Minato[2]

1. U-Tokyo
2. Hokkaido Univ.
3. VOYAGE GROUP Inc.

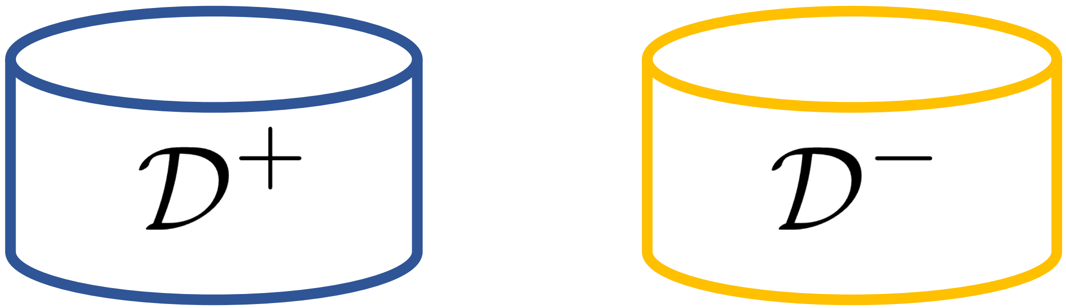
(to appear in KDD2017 research track)

Single-page Summary

- We study **Emerging Pattern Mining** (EPM) with **statistical guarantee**.
  - EPM is also known as *contrast set mining*, *subgroup mining.*
- We propose **two-stage mining methods** that control **FWER** or **FDR**.
  - FWER: Family-Wise Error Rate
  - FDR: False Discovery Rate

Emerging Pattern Mining

- Items: $I = \{1, \ldots, |I|\}$
- Pattern: $e \subseteq I$
- Emerging Pattern:
  - e that appears frequently in D^+ but not in D^-.



$\mathcal{D}^+$	$\mathcal{D}^-$
{**1, 2**}	{1, 3}
{1, 3, 4}	{2, 4}
{**1, 2**, 3}	{1, 3, 4}
…	…

Emerging Pattern Mining

- Database:
  - $D = \{ (x_i, y_i) : i = 1,\ldots,N \}$ where $x_i \subseteq I, y_i \in \{0,1\}$
  - $D^+ = \{ (x, y) \in D : y = 1 \}$
  - $D^- = \{ (x, y) \in D : y = 0\}$
- Emerging Pattern: e s.t. $N_e^+ / N_e^- > a$
  - $N_e^+ = |\{ (x, y) \in D^+ : e \subseteq x \}|$
  - $N_e^- = |\{ (x, y) \in D^- : e \subseteq x \}|$
  - $a =$ a given threshold
- Problems of Emerging Pattern Mining:

1. Too many *insignificant* patterns are found.
2. Not sure whether the found patterns are just *random fluctuation* of D or truly significant.

Statistical Emerging Pattern Mining

- Assumption: (x, y) ~i.i.d.~ $\mathbb{P}[x, y]$
  - $\mathbb{P}[x, y]$ = *unknown* true distribution
- Positive label prob.: $\mu_e = \mathbb{P}[y = 1 \mid e \subseteq x]$
- True / False Emerging Pattern:
  - $\varepsilon_{\mathrm{true}} = \{e \in 2^I : \mu_e > a\}$
  - $\varepsilon_{\mathrm{false}} = \{e \in 2^I : \mu_e \leqq a\}$
- SEMP: Estimate $\varepsilon_{\mathrm{true}}$ from D.

Statistical Emerging Pattern Mining

- ε_{alg} = outputs of an algorithm
- Family-wise error rate (FWER):
  - $\text{FWER} = \mathbb{P}[\, |\varepsilon_{\text{alg}} \cap \varepsilon_{\text{false}}| \geqq 1 \,]$
- False discovery rate (FDR):
  - $\text{FDR} = \mathbb{E}[\, |\varepsilon_{\text{alg}} \cap \varepsilon_{\text{false}}| \,/\, |\varepsilon_{\text{alg}}| \,]$

We propose two-stage mining methods that satisfy $\text{FWER} \leqq q$ or $\text{FDR} \leqq q$.

Pattern as a hypothesis

- Null hypothesis ($e \in \varepsilon_{\text{false}}$):
  - $H_e^0 : \mu_e = a$
- Alternative hypothesis ($e \in \varepsilon_{\text{true}}$):
  - $H_e^1 : \mu_e > a$
- P-value:

$$p_e = \mathbb{P}\left[\mathrm{Sup}\left(e; \mathcal{D}^+\right) \geq N_e^+ \mid \mathrm{Sup}(e; \mathcal{D}) = N_e, H_e^0\right]$$

$$= \sum_{n=N_e^+}^{N_e} \binom{N_e}{n} a^n (1-a)^{N_e - n}.$$

Multiple Testing Correction

- P-value:
  - How data is likely to be generated under H_e^0.
  - Small p-value → Rare event
- Single Hypothesis:
  - Reject e if $p_e \leqq q$ ➔ We think e is a True EP.
  - Control FWER (and FDR) $\leqq q$.
- Multiple hypotheses:
  - Probability of including "*false positive*" gets fairly high.
  - Peeking p-values causes "*selection bias*".

Need multiple testing correction.

Bonferroni correction for FWER

- Reject e if $p_e \leqq q / m$.
  - m = # of patterns to test.
- Control FWER at level q.

Step-up correction for FDR

- Reject $e_{(1)},\ldots,e_{(k)}$ $\quad p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$
  - $p_{(i)}$ = the i-th smaller p-value
  - $e_{(i)}$ = its corresponding pattern
  - $k = \arg\max_{0 \leq i \leq m} \left\{ p_{(i)} \leq \frac{q}{c(m)} \frac{i}{m} \right\}$ ← Step-up rejection Level
  - BH : $c(m) = 1$
  - BY : $c(m) = \Sigma_{i=1}^{m} (1 / i)$
- Control FDR at level q,
  - under independence among hypotheses (BH)
  - or under arbitrary correlations (BY).

But patterns are exponentially large…

- m (= # of patterns to test) can be exponentially large : $2^{|I|}$.

- Naïve Bonferroni / Step-up corrections cannot find much patterns.
  - Both corrections have q / m factor

Needs to reduce # of patterns to test.

Existing statistical pattern mining and our result

	Existing	Our result: Two-stage Mining methods	
	LAMP	**LAMP-EP**	**QT-LAMP-EP**
Mining target	SAM	SEPM	SEPM
Multiple Testing	FWER	FWER	FDR
Pattern Reduction	Testable	Testable	Quasi-Testable
Testing method	Bonferroni	Bonferroni	Step-up

- LAMP [Terada+ 13]
  - Proposed for statistical association mining (SAM)
  - Selects testable patterns before correction, and
  - Controls FWER.

Two-stage Mining Method

1. Find a “appropriate” threshold τ.
2. Test patterns $\{e : N_e > \tau\}$ with multiple testing correction.

How to choose “appropriate” τ?
We use “testability” just like LAMP!

Testability for FWER

- Tarone's exclusion principle:
  - Patterns with large p-value can be omitted without testing; it cannot be significant.
  - FWER can be controlled by $q \,/\, m_{\text{test}}$ (not m).
    - m_{test} = # of "testable" patterns

- $\psi(N_e)$ = a lower bound of p_e
  - $\psi(N) = a^N$

- e is testable if $\psi(p_e) \leqq q \,/\, m_{\text{test}}$

LAMP-EP (LAMP for SEPM)

- Finding the largest testable set boils down to finding the following threshold τ_{FWER} s.t.:

$$\psi(\tau_{\mathrm{FWER}} - 1) > \delta_{\mathrm{FWER}}(\tau_{\mathrm{FWER}} - 1; q, \mathcal{D}), \qquad \ldots(1)$$

$$\psi(\tau_{\mathrm{FWER}}) \leq \delta_{\mathrm{FWER}}(\tau_{\mathrm{FWER}}; q, \mathcal{D}), \text{ where} \qquad \ldots(2)$$

$$\delta_{\mathrm{FWER}}(\tau; q, \mathcal{D}) = \frac{q}{|\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D})|}, \quad (=\text{Tarone})$$

$\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D})$: Frequent patterns with min-support τ.

(1)…patterns of support $<$ τ_{FWER} are untestable.
(2)…patterns of support $\geq$ τ_{FWER} are testable.

Next : Controlling FDR

- **No** principled method yet.
- Major challenges:

1. No Tarone's exclusion in FDR:
   - We solve this by splitting the dataset into calibration and main datasets.
2. Not sure how to select a "testable" set.
   - We introduce "quasi-testablity" (approximated testability).

Quasi-testability for FDR

- Step-up Correction for FDR:
  - Reject $e_{(1)},\ldots,e_{(k)}$
    - $p_{(i)}$ = the i-th smaller p-value. $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$.
    - $e_{(i)}$ = its corresponding pattern.
    - $k = \arg\max_{0 \leq i \leq m} \left\{ p_{(i)} \leq \frac{q}{c(m)} \frac{i}{m} \right\}$
- e is testable if $\psi(N_e) \leqq p_{(k)}$
  - But $p_{(k)}$ must be unknown to avoid selection bias.
- e is quasi-testable if $\psi(N_e) \leqq p_{\mathrm{est}}$
  - Instead of true $p_{(k)}$, we use an estimator p_{est}.
  - We split D into D_{main} and D_{carib}, and use D_{carib} to estimate p_{est}.

QT-LAMP-EP for controlling FDR

- Find threshold value τ_{FDR} such that

$$\psi(\tau_{\mathrm{FDR}}-1) > \delta_{\mathrm{FDR}}(\tau_{\mathrm{FDR}}-1; q, \mathcal{D}_{\mathrm{carib}}) \qquad \ldots(1)$$

$$\psi(\tau_{\mathrm{FDR}}) \leq \delta_{\mathrm{FDR}}(\tau_{\mathrm{FDR}}; q, \mathcal{D}_{\mathrm{carib}}) \;\;, \text{ where } \qquad \ldots(2)$$

$$\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}}) = \frac{q}{c(|\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{carib}})|)} \frac{\hat{k}(\tau; \mathcal{D}_{\mathrm{carib}})}{|\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{carib}})|} \quad \ldots(3)$$

$\hat{k}(\tau; \mathcal{D}_{\mathrm{carib}})$ = # of patterns rejected if step-up method is conducted for $\mathcal{D}_{\mathrm{carib}}$.

(1)…patterns of support < τ_{FDR} are untestable,
(2)…patterns of support ≥ τ_{FDR} are testable,
under estimated step-up rejection level of (3).

Computer Simulations

- Statistical powers of LAMP-EP and QT-LAMP-EP are compared.
- LAMP-EP used the entire dataset for testing.
- QT-LAMP-EP used 20% of D as D_{carib} for obtaining p_{est}, and 80% as D_{main} for testing.

FWER/FDR in synthetic dataset.

- Synthetic patterns, $a = 0.5, q = 0.05$.
- True SEPs: subset of $\{1, \dots, 10\}$: $\mu_e = 0.7$
- Other patterns: subset of $\{11, \dots, 100\}$, $\mu_e = 0.5$.
- $|\mathcal{D}| = 10^5$, 10% of patterns are true SEPs.

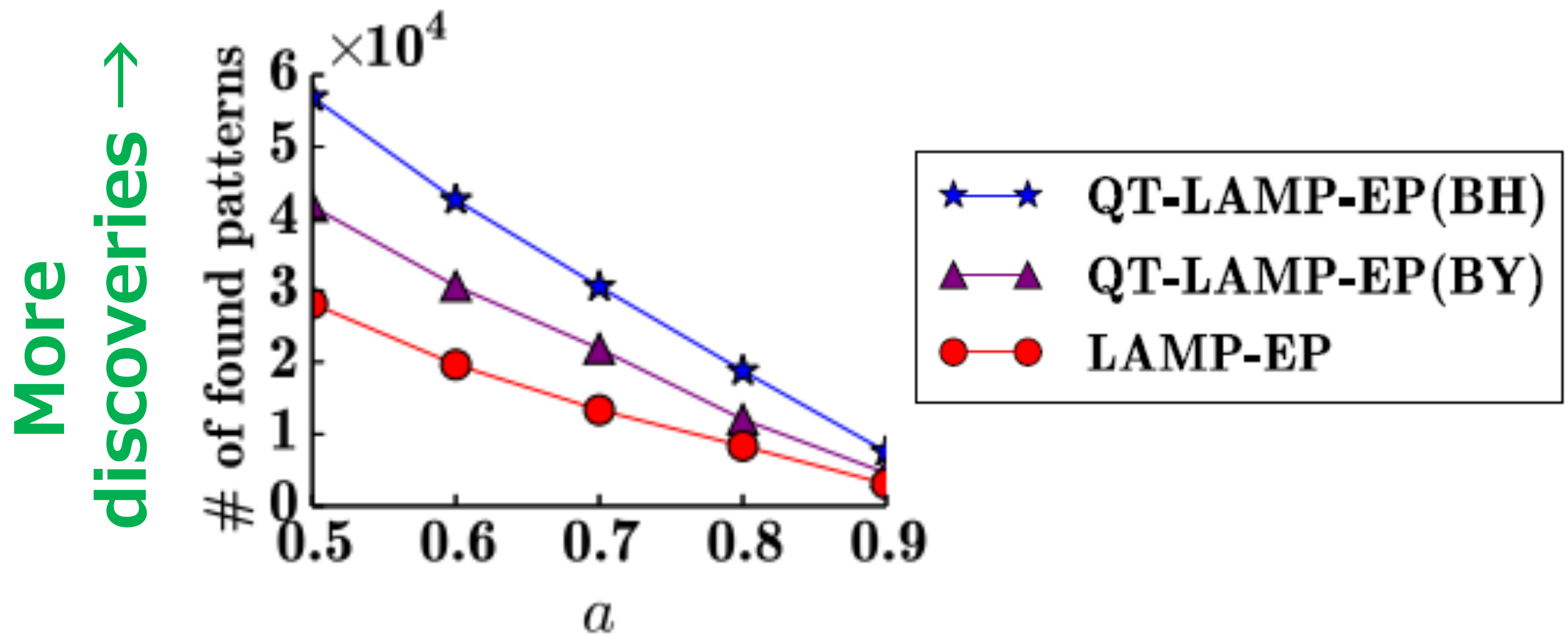
Results:

algorithms	# of TDs	# of FDs	FDR	FWER
EPM	516.50	3848.61	0.88	1.00
LAMP-EP	166.32	0.01	$\boxed{6.02e-05}$	$\boxed{0.01}$
QT-LAMP-EP (BH)	230.87	4.10	$\boxed{0.017}$	0.99
QT-LAMP-EP (BY)	184.10	0.40	$\boxed{2.13e-03}$	0.32

Controlled $\leqq q$

Number of discoveries in a classification (mushroom) dataset

- Controlling FDR yields more patterns than FWER.



- Similar results hold for other 7 datasets.

Conclusion

- We formulated **statistical emerging pattern mining**.
- We propose **two-stage mining methods**:
  - LAMP-EP controls **FWER**.
  - QT-LAMP-EP controls **FDR**.
- We empirically verified their statistical power.

Thanks!

Contact us

Paper and software are available.

Paper:
http://www.tkl.iis.u-tokyo.ac.jp/~jkomiyama/pdf/kdd-statistical-emerging.pdf

Software:
https://github.com/jkomiyama/qtlamp

Contact:
Junpei Komiyama
junpei@komiyama.info