# Controlling FWER and FDR in Emerging Pattern Mining

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.

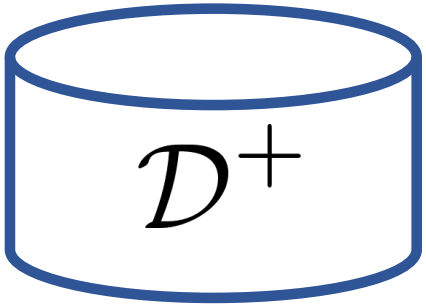
東京大学
生産技術研究所
Institute of Industrial Science,
The University of Tokyo

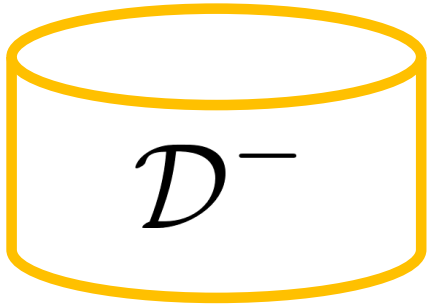
# Single-page Summary

- We study **Emerging Pattern Mining** (aka. contrast set mining, subgroup mining) with statistical guarantee.
  - Pattern = itemset $I = \{1, 2, \ldots, \ell\}$.
- We propose a two-stage pattern mining method that controls not only **FWER**, but **FDR**.

# Emerging pattern mining (EPM)

- EP: pattern that appears frequently in a dataset but not in the other dataset.




{**1, 2**}
{1, 3, 4}
{**1, 2**, 3}
…

{1, 3}
{2, 4}
{1, 3, 4}
…

Applications: binary classification, feature selection, change point detection, etc.

# Emerging pattern mining (EPM) cont.

- $\mathcal{D}^+, \mathcal{D}^-$ :appearance of pattern $e$ in corresponding dataset.
- Standard objective of EPM: Enumerate all pattern $e$ s.t. $N_e^+ / N_e^- > a$ for given threshold $a$, where $N_e^+$ and $N_e^-$ are supports (# of occurrences) of pattern $e$ in $\mathcal{D}^+, \mathcal{D}^-$.
- Problems:

1. Too many insignificant patterns (most with small supports $N^+, N^-$) are found.
2. Not sure whether the found pattern are just random fluctuation of datasets or truly significant.

# Statisical Emerging Pattern Mining (SEPM)

We formalize statistical EPM as follows.

- Let a datapoint is i.i.d. sample from some distribution $\mathbb{P}[x, y]$.
  - $x \subseteq I$ : itemset, $y \in \{0, 1\}$ : label.
- Let $\mathcal{D} = \{(x_i, y_i)\} = \mathcal{D}^+ \cup \mathcal{D}^-$, where

$$\mathcal{D}^+ = \{(x_i, y_i) \in \mathcal{D}, y = 1\}$$

$$\mathcal{D}^- = \{(x_i, y_i) \in \mathcal{D}, y = 0\}.$$

- Statistical EPM: find pattern $e$ with positive label probability $a > 0$:

$$\mathbb{P}[y = 1 \mid e \subseteq x] > a.$$

# Statisical Emerging pattern mining (SEPM)

- True and false SEP: let $\mu_e = \mathbb{P}[y = 1 \mid e \subseteq x]$ and

$$\mathcal{E}_{\text{true}} = \{e \in 2^I : \mu_e > a\}$$

$$\mathcal{E}_{\text{false}} = \{e \in 2^I : \mu_e \leq a\}.$$

- Let $\mathcal{E}_{\text{alg}} \subseteq 2^I$ be the output of a pattern mining algorithm.
- An algorithm controls FWER with level $q$ if

$$\mathbb{P}[|\mathcal{E}_{\text{alg}} \cap \mathcal{E}_{\text{false}}| \geq 1] \leq q.$$

- An algorithm controls FDR with level $q$ if

$$\mathbb{E}\left[\frac{|\mathcal{E}_{\text{alg}} \cap \mathcal{E}_{\text{false}}|}{|\mathcal{E}_{\text{alg}}|}\right] \leq q.$$

# Pattern as a hypothesis

- Null hypothesis

$$H_e^0 : \mu_e = a.$$

- Alternative hypothesis

$$H_e^1 : \mu_e > a.$$

  - Pattern $e$ is true SEP.

- P-value:

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

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

# Concentration inequality

- $P_e$ for large $N_e$ is approximated by the Chernoff concentration inequality:

$$p \approx p_e^{\mathrm{C}} \leq \begin{cases} \exp(-N_e d_{\mathrm{KL}}(\hat{\mu}_e, a)) & (\hat{\mu}_e > a), \\ 1 & (\text{otherwise}), \end{cases}$$

, where

$$d_{\mathrm{KL}}(p, q) := p \log (p/q) + (1-p) \log((1-p)/(1-q))$$

(Bernoulli KL divergence).

# Frequent pattern mining (FPM)

- Given dataset $\mathcal{D}$ and minimum support $\tau$, the goal of FPM is to enumerate all pattern $e \in 2^I$ of its support $\tau$ or larger.
- LCM [Uno+ 03] is one of the fastest FPM algorithms.
- We use LCM as an subroutine for SEPM.

# Bonferroni correction for FWER

- Let $m$ be # of patterns to test. Then,
- Reject each hypothesis (= pattern) with p-value threshold $p_e \leq q/m$.
  - Controls FWER at level $q$.

# Step-up methods for FDR

- Sort patterns by their p-values as

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}.$$

- Let $k = \arg\max_{0 \leq i \leq m} \left\{ p_{(i)} \leq \frac{q}{c(m)} \frac{i}{m} \right\}$.
- Adjustment $c(m) = 1$ (Benjamini-Hochberg, BH), $c(m) = \sum_{i=1}^{m} (1/i)$ (Benjamini-Yekutieli, BY)
- BH / BY control FDR at level $q$, under independence among hypotheses (BH), under arbitrary correlations (BY).

# Patterns are exponentially large

- # of possible patterns ($= 2^{|I|}$) is exponentially large to # of items.
- Large set of hypotheses -> Weak statistical power of multiple hypotheses.
- Needs to reduce # of patterns to test.
  - LAMP [Terada+ 13]: only “testable” patterns

# LAMP [Terada et al. 2013]

- A pattern (itemset) mining algorithm for controlling FWER in statistical association.

| | 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 |

# Minimum p-value and testability

- P-value:

$$p_e = \mathbb{P}\left[\mathrm{Sup}\left(e;\mathcal{D}^+\right) \geq N_e^+ \mid \mathrm{Sup}(e;\mathcal{D}) = N_e, H_e^0\right]$$
$$= \sum_{n=N_e^+}^{N_e} \binom{N_e}{n} a^n (1-a)^{N_e - n}.$$

- A pattern e is testable with threshold q if is possible p-value is less than q:

$$p_e = \min_{N_e^+ \leq N_e} p_e(N_e^+)$$

# Maximize statistical power in FWER

- LAMP [Terada+ 13] controls FWER by using the following Tarone's exclusion principle:
- A hypothesis is testable if p_e = ¥min_{N_e^-} p_e(N_e) is larger than q/m.
- Corrections factor m can be reduced to the number of testable hypotheses:
- In our EPM: "see unlabeled dataset D=[x]". If the p-value of pattern e cannot be below q/m with arbitrary labelling y, then remove the pattern e without testing it."

# LAMP-EP (LAMP for EP mining)

- LAMP “finds a largest set of testable patterns and conducts FWER testing”.
- Find the largest set boils down to finding the following threshold $\tau^*_{\mathrm{FWER}} \in \mathbb{N}$ such that:

$$\psi(\tau^*_{\mathrm{FWER}} - 1) > \delta_{\mathrm{FWER}}(\tau^*_{\mathrm{FWER}} - 1; q, \mathcal{D}),$$
$$\psi(\tau^*_{\mathrm{FWER}}) \leq \delta_{\mathrm{FWER}}(\tau^*_{\mathrm{FWER}}; q, \mathcal{D}),$$

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

# Controlling FDR in pattern mining tasks

- No principled method yet.
- Major challenges:
  - No Tarone's principle in FDR:
    - Gilbert's exclusion for FDR [] requires independence among hypotheses, but patterns are highly correlated.
    - -> Solve this problem by dividing calibration/main dataset
  - Not sure how to select a "testable" set.
    - -> Solve this by introducing "quasi-testable" set.

# QT-LAMP-EP (controlling FDR in EPM)

- Similar to LAMP-EP, we first choose $\tau_{\mathrm{FDR}}$ and test patterns with its support $\tau_{\mathrm{FDR}}$ or larger
- Instead of Tarone's principle, we split datasets into calibration dataset $\mathcal{D}_{\mathrm{carib}}$ and main dataset $\mathcal{D}_{\mathrm{main}}$.
  - The caribration dataset is for determining $\tau_{\mathrm{FDR}}$, and mulitiple testing (step-up method) is conducted for the main dataset.

# Quasi-testable set

- "patterns with support smaller than $\tau_{\mathrm{FDR}}$ cannot be rejected for any labelling of $y$".

- Find $\tau_{\mathrm{FDR}}$ such that

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

Where

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

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

# QT-LAMP-EP summary

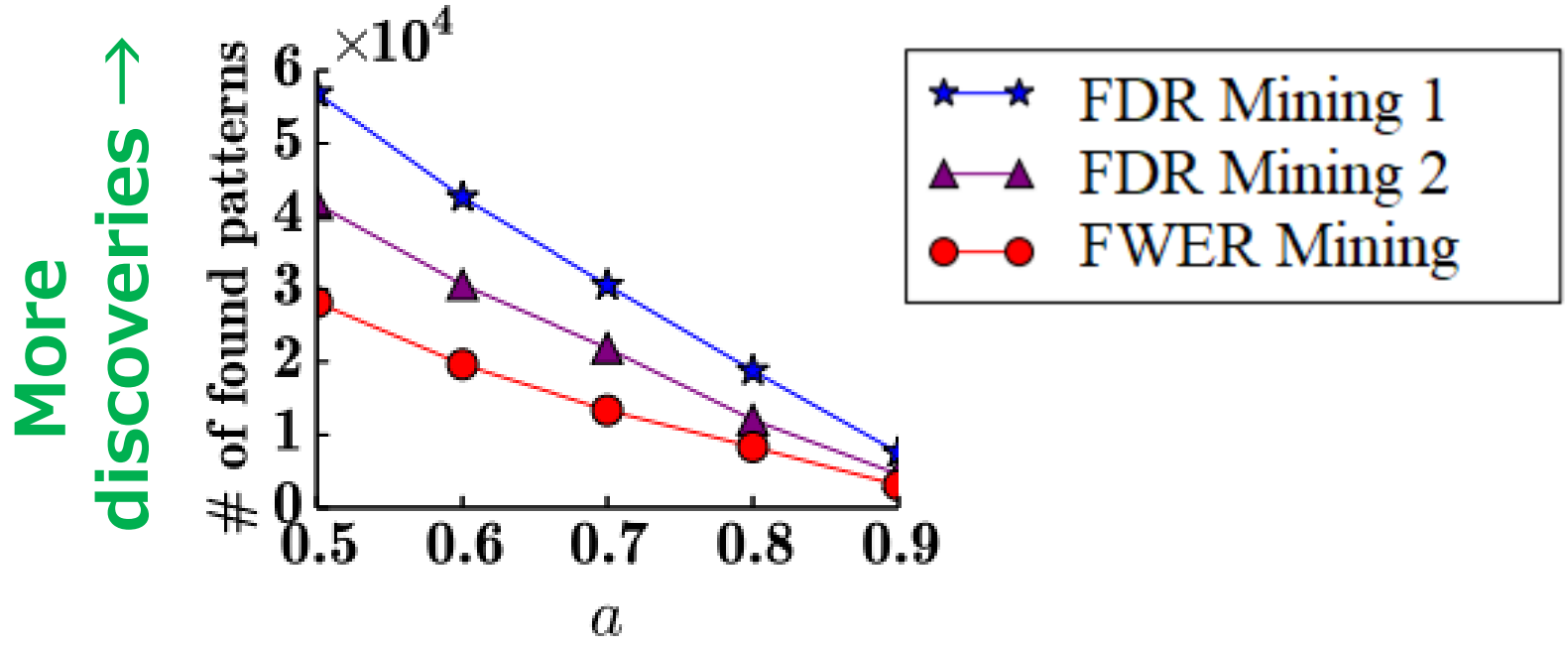
- Find $\tau_{\mathrm{FDR}}$ by bisection search over $1, \ldots, |\mathcal{D}|$ by using calibration dataset.
- Conduct step-up methods for all patterns with its support $\tau_{\mathrm{FDR}}$ or larger by using main dataset.

# Experiment

- Statistical powers of LAMP-EP and QT-LAMP-EP are compared.
- Binary classification datasets are used.
- Similar results for other datasets, value of a.
- Carib/main datasetの設定

# Simulation: FDR v.s. FWER

- Using FDR yields more patterns than FWER!



# ここから先は不使用

# Two-stage framework for controlling FWER/FDR (old)

- Selection threshold $\tau$: determine the set of patterns to test.
  - Proper choice of $\tau$ maximizes the number of found patterns.

1. find proper value of $\tau$ such that patterns with a support $\tau$ or larger can be significant.
2. Conduct standard multiple testing with patterns of supports $\tau$ or larger.

# Contact us

Paper and software are available.

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

Contact:
Junpei Komiyama
junpei@komiyama.info