Statistical Emerging Pattern Mining with Multiple Testing Correction

Junpei Komiyama
The University of Tokyo
junpei@komiyama.info

Masakazu Ishihata
Hokkaido University
ishihata.masakazu@ist.hokudai.ac.jp

Hiroki Arimura
Hokkaido University
arim@ist.hokudai.ac.jp

Takashi Nishibayashi
VOYAGE GROUP, Inc.
takashi.nishibayashi@gmail.com

Shin-ichi Minato
Hokkaido University
minato@ist.hokudai.ac.jp

ABSTRACT

Emerging patterns are patterns whose support significantly differs between two databases. We study the problem of listing emerging patterns with a multiple testing guarantee. Recently, Terada et al. proposed the Limitless Arity Multiple-testing Procedure (LAMP) that controls the family-wise error rate (FWER) in statistical association mining. LAMP reduces the number of "untestable" hypotheses without compromising its statistical power. Still, FWER is restrictive, and as a result, its statistical power is inherently unsatisfying when the number of patterns is large. On the other hand, the false discovery rate (FDR) is less restrictive than FWER, and thus controlling FDR yields a larger number of significant patterns. We propose two emerging pattern mining methods: the first one controls FWER, and the second one controls FDR. The effectiveness of the methods is verified in computer simulations with real-world datasets.

KEYWORDS

emerging pattern mining, multiple testing, statistical pattern mining

1 INTRODUCTION

Finding differences between two datasets is of fundamental importance in many scientific fields. Many tasks, such as binary classification, feature selection, change point detection, and concept drift learning, boil down to finding good discriminative features that explain the differences. When the structure of the problem is simple, it suffices to consider each single feature separately. However, when the problem is not straightforward, the combinatorial effect of multiple features can be fundamentally important. Since the number of possible combinations is exponential to the number of features, this task is inherently challenging.

Emerging pattern mining (EPM) [10] is an approach that lists the patterns where the difference between two datasets is larger than a given threshold. One of the greatest advantages of EPM lies in that it can naturally find combinatorial features: each combinatorial feature corresponds to an itemset in pattern mining. Although there are many studies on EPM (for a detailed list of these papers, see Dong and Bailey [9]. Moreover, for similar concepts such as contrast set mining and subgroup mining, see Novak et al. [22]), the standard formulation of EPM lacks a statistical assessment, which induces a risk that a significant fraction of found patterns are just false positives; that is, the patterns may have only been found because of the random nature of data and are actually insignificant. Statistical testing is widely used to control such risks. In particular, a procedure of listing patterns can be verified within the framework of multiple hypothesis testing by considering each itemset to be a hypothesis.

Undoubtedly, we would like to find as many patterns as possible for a given level of risk. The study of multiple testing is somewhat paradoxical: the larger the number of patterns to test is, the less powerful the test becomes. This is because that the risk of finding false discoveries rises as the number of patterns to test increases. Since the number of possible patterns is 2^ℓ to the number of the item ℓ, testing all the patterns is unlikely to result in many patterns. Moreover, testing all possible patterns is computationally prohibitive. In order to avoid this "combinatorial explosion curse", we would like to determine a set of patterns to test before conducting multiple testing. Unfortunately, most of the existing data mining approaches that take multiple testing into consideration are not principled when it comes to determining the patterns to test: they choose the minimum support by using implicit knowledge. If one selects a set of hypotheses by looking into the database, that can cause a selection bias that devastates the entire testing process. A recently proposed method called Limitless-Arity Multiple-testing Procedure (LAMP) [28] avoids this pitfall and enables us to select the patterns to test in a principled way; it effectively reduces the number of patterns without prior knowledge and alleviates the combinatorial explosion curse.

There are two different targets to control in multiple testing. Namely, the family-wise error rate (FWER) and the false discovery rate (FDR) [4]. The former value is the probability of a false discovery among the found patterns. The aforementioned LAMP is designed to control FWER. Although control of FWER is crucial in some areas such as genome association studies, where a false discovery can cause severe harm, the restrictive nature of FWER often limits the statistical power. We hence argue that, in many cases, some number of false discoveries is tolerable. For example, feature selection is not harmed much when insignificant features are added. Moreover, E-commerce data is often analyzed with the hope of

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD'17, August 13–17, 2017, Halifax, NS, Canada.
© 2017 Copyright held by the owner/author(s). Publication rights licensed to ACM.
978-1-4503-4887-4/17/08…$15.00
DOI: http://dx.doi.org/10.1145/3097983.3098137

making discoveries that can boost revenue; a certain number of false discoveries would be tolerable in this case as well. For such applications, a more modern notion of FDR, i.e., is the probability of false discoveries among all discoveries, is less restrictive and would empower one to make discoveries. Unfortunately, there is as yet no pattern mining method for controlling FDR with an ability to choose an appropriate set of patterns to test.

The statistical pattern mining is essentially comprised of two stages: the first stage selects a set of patterns to test, and the second stage tests the selected patterns. While LAMP is dependent on Tarone's exclusion principle [26] that excludes "untestable" patterns, there is no corresponding notion in FDR, and thus, one must devise a principled way to reduce a set of patterns. In this paper, we describe a way to control FDR (Table 1). The key notion for testing FDR is "quasi-testability". The quasi-testability takes the adaptive nature of FDR into consideration, and hence, it effectively reduces the hypotheses that are very unlikely to be significant.

The contributions of this paper are as follows. We formalize the problem of statistical emerging pattern mining (SEPM). We propose LAMP-EP, a version of LAMP for SEPM, which effectively lists patterns while controlling FWER. Moreover, we propose QT-LAMP-EP which controls FDR in SEPM. Note that the quasi-testability based method is not only for emerging patterns and possibly be applied to any pattern mining method. To verify the performance of the methods, computer simulations were conducted with eight real-world datasets. The exhaustive set of simulations shows that tolerance to a predefined FDR can lead to a significantly larger set of discoveries.

2 RELATED WORK

Some studies are based on distributional assumptions. Kirsch et al. [14] developed a multiple testing procedure that identifies the best support threshold such that the pattern will be significant by assuming the supports of patterns are Poisson-distributed. Low-Kam et al. [18] proposed a mining method that measures the significance of a pattern by the deviation from the null model. In this paper, we do not assume such models; the only essential assumption we rely on is the i.i.d. property of the transactions.

Unlike most of the existing work, the LAMP algorithm [28] does not require a stringent distributional assumption and is capable of determining the minimum support threshold. Some subsequent studies were inspired by LAMP: Sugiyama et al. [25] considered statistical association in subgraph mining. Terada et al. [27] and Felipe et al. [17] integrate the Westfall-Young permutation method [32] into the pattern selection and showed that it empirically yields more discoveries than LAMP. Note that the Westfall-Young method exploits subset pivotality of SAM [20] and not applicable to EPM. Papaxanthos et al. [23] extended LAMP to cases where categorical covariates exist. Webb et al. [31] studied association rule mining and proposed two methods to control the error rate. One method uses explanatory and holdout datasets: the explanatory dataset is for pattern discovery and the holdout dataset is for pattern testing. Although this method is only for association rule mining, our method is partly inspired by theirs in that it uses several independent datasets. Riondato and Vandin [24] studied the association rule mining. They conducted a statistical learning theory based analysis that is distribution-free and derived a guarantee that is similar to FWER. Their analysis requires a minimum support threshold. Hanhijärvi [12] proposed a procedure to adjust the Bonferroni correction factor to control FWER by using randomization under the assumption of subset pivotality. Lallich et al. [15, 16] proposed a bootstrap-based algorithm for association mining that controls a version of FWER that tolerates a fixed number of false discoveries.

Table 1: Comparison of methods for controlling FWER and FDR in pattern mining. LAMP is a statistical association mining (SAM) method that controls FWER by applying the Bonferroni method [7] to testable patterns. This paper proposes LAMP-EP and QT-LAMP-EP that control FWER and FDR in the SEPM problem. QT-LAMP-EP applies the step-up method [6, 13] to quasi-testable patterns.

	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

3 PROBLEM SETUP

In this section, we first briefly review the *frequent pattern mining* (FPM) and *emerging pattern mining* (EPM) problems. Then, we propose a new problem called *statistical emerging pattern mining* (SEPM), which is an extension of EPM that assesses the chance of a discovered pattern being a true discovery.

3.1 Frequent Pattern Mining (FPM)

Let $\mathcal{D}$ be a database, which is a set of transactions. Each transaction is a tuple (x, y), where x is a *pattern* and $y \in \{0, 1\}$ is a *label* of pattern x. Each pattern x is a subset of ℓ items indexed as $I = \{1, 2, \ldots, \ell\}$; namely, $x \in 2^I$. A pattern with a label 0 (resp. 1) is said to be *negative* (resp. *positive*). Given a dataset $\mathcal{D}$, we use $\mathcal{D}^+$ and $\mathcal{D}^-$ to denote sub-datasets of $\mathcal{D}$ that consist of positive and negative patterns; namely,

$$\mathcal{D}^+ = \{(x, y) \in \mathcal{D} : y = 1\}, \tag{1}$$
$$\mathcal{D}^- = \{(x, y) \in \mathcal{D} : y = 0\}. \tag{2}$$

We assume that each transaction $(x, y) \in \mathcal{D}$ is an i.i.d. sample from an unknown joint distribution $\mathbb{P}[x, y]$ on $2^I \times \{0, 1\}$; namely, $\mathcal{D}$ is a random variable. All probability and expectations in this paper are taken on this distribution.

Given a database $\mathcal{D}$ and a pattern $e \in 2^I$, the *occurrences* of e comprise a set of transactions in $\mathcal{D}$ that contains e. The *support* of e is the number of occurrences of e. We use $\mathrm{Occ}(e; \mathcal{D})$ and $\mathrm{Sup}(e; \mathcal{D})$ to denote the occurrences and the support of e:

$$\mathrm{Occ}(e; \mathcal{D}) = \{(x, y) \in \mathcal{D} : e \subseteq x\}, \tag{3}$$
$$\mathrm{Sup}(e; \mathcal{D}) = |\mathrm{Occ}(e; \mathcal{D})|. \tag{4}$$

Given a dataset $\mathcal{D}$ and a minimum support τ ($\tau = 0, 1, \ldots, |\mathcal{D}|$), a *frequent pattern* (FP) is a pattern that appears τ or more times in

$\mathcal{D}$. We use $\mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D})$ to denote the set of all frequent patterns:

$$\mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D}) = \{e \in 2^I : \mathrm{Sup}(e;\mathcal{D}) \geq \tau\}. \tag{5}$$

The goal of the *frequent pattern mining* (FPM) [1, 3] is to find $\mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D})$ given τ and $\mathcal{D}$. An item $i \in I$ is said to be *redundant* in $\mathcal{D}$ if $\mathrm{Occ}(\{i\};\mathcal{D}) = \mathcal{D}$; in other words, i appears in all transactions in $\mathcal{D}$. In this paper, we assume that the given dataset $\mathcal{D}$ contains no redundant item. Consequently, the only pattern $e \in 2^I$ that satisfies $\mathrm{Occ}(e;\mathcal{D}) = \mathcal{D}$ is the empty pattern $e = \phi$; namely, $\mathcal{E}_{\mathrm{FP}}(|\mathcal{D}|;\mathcal{D}) = \{\phi\}$.

3.2 Emerging Pattern Mining (EPM)

Whereas FPM does not consider the labels of patterns, the goal of *emerging pattern mining* (EPM) [10] is to list all patterns which frequently appear in $\mathcal{D}^+$ but not in $\mathcal{D}^-$. Let $\mathrm{GR}(e;\mathcal{D})$ be the *growth rate* of pattern e that is defined as follows:

$$\mathrm{GR}(e;\mathcal{D}) = \begin{cases} 0 & (\mathrm{Sup}(e;\mathcal{D}^+) = 0 \text{ and } \mathrm{Sup}(e;\mathcal{D}^-) = 0), \\ \infty & (\mathrm{Sup}(e;\mathcal{D}^+) \neq 0 \text{ and } \mathrm{Sup}(e;\mathcal{D}^-) = 0), \\ \frac{\mathrm{Sup}(e;\mathcal{D}^+)}{\mathrm{Sup}(e;\mathcal{D}^-)} & (\text{otherwise}). \end{cases} \tag{6}$$

Given a growth rate threshold $a > 0$, *emerging patterns* (EPs) are defined as follows:

$$\mathcal{E}_{\mathrm{EP}}(a;\mathcal{D}) = \left\{e \in 2^I : \mathrm{GR}(e;\mathcal{D}) > a\right\}. \tag{7}$$

The EPM problem is to find $\mathcal{E}_{\mathrm{EP}}(a;\mathcal{D})$ given a and $\mathcal{D}$. Because listing the entire $\mathcal{E}_{\mathrm{EP}}(a;\mathcal{D})$ is often computationally prohibitive, many algorithms for EPM [2, 10, 19] set a minimum support τ for both $\mathrm{Sup}(e;\mathcal{D}^+)$ and $\mathrm{Sup}(e;\mathcal{D}^-)$; in other words, usually only patterns $e \in \mathcal{E}_{\mathrm{EP}}(a;\mathcal{D})$ such that $\mathrm{Sup}(e;\mathcal{D}^+) \geq \tau$ and $\mathrm{Sup}(e;\mathcal{D}^+) \geq \tau$ are enumerated.

The above formulation lacks a statistical assessment: for example, how many of the found EPs are reproducible when we conduct an EPM with another dataset $\mathcal{D}'$ generated by the same process? In the next section, we extend EPM to statistically assess the found patterns.

3.3 Statistical Emerging Pattern Mining (SEPM)

We first define the true emerging patterns independently of the observed dataset $\mathcal{D}$. Let μ_e be the *positive label probability* of pattern $e \in 2^I$ that is defined as follows:

$$\mu_e = \mathbb{P}[y = 1 \mid e \subseteq x] \propto \sum_{e' \in 2^I : e \subseteq e'} \mathbb{P}\left[y = 1, x = e'\right]. \tag{8}$$

Given a *positive label probability threshold* $a \in (0,1)$, we define the *true emerging patterns* $\mathcal{E}_{\mathrm{true}}$ and *false patterns* $\mathcal{E}_{\mathrm{false}}$ as follows:

$$\mathcal{E}_{\mathrm{true}} = \{e \in 2^I : \mu_e > a\}, \tag{9}$$

$$\mathcal{E}_{\mathrm{false}} = \{e \in 2^I : \mu_e \leq a\}. \tag{10}$$

Since μ_e is unobservable, we need to estimate whether or not each pattern e lies within $\mathcal{E}_{\mathrm{true}}$ from the observed dataset $\mathcal{D}$.

Here, we describe a new problem, called *statistical emerging pattern mining* (SEPM), which is to find a fraction of the true emerging patterns $\mathcal{E}_{\mathrm{true}}$ from the given dataset $\mathcal{D}$ with a statistical error bound. Ultimately, we would like to construct an algorithm that exactly finds $\mathcal{E}_{\mathrm{true}}$; however, this is impossible because of the random nature of $\mathcal{D}$. Instead of finding the whole $\mathcal{E}_{\mathrm{true}}$, we attempt to construct an algorithm that finds a set of appropriate patterns $\mathcal{E}_{\mathrm{alg}} \subseteq 2^I$ and keeps its error rate under the given significance level $q \in (0,1)$. We consider two types of error rate, the *family-wise error rate* (FWER) and the *false discovery rate* (FDR), defined as follows:

Definition 3.1 (FWER). Given a significance level q, an SEPM algorithm is said to control FWER if

$$\mathbb{P}\left[|\mathcal{E}_{\mathrm{alg}} \cap \mathcal{E}_{\mathrm{false}}| \geq 1\right] \leq q. \tag{11}$$

FWER is the probability that $\mathcal{E}_{\mathrm{alg}}$ contains a false pattern $e \in \mathcal{E}_{\mathrm{false}}$.

Definition 3.2 (FDR). Given a significance level q, an SEPM algorithm is said to control FDR if

$$\mathbb{E}\left[\frac{|\mathcal{E}_{\mathrm{alg}} \cap \mathcal{E}_{\mathrm{false}}|}{|\mathcal{E}_{\mathrm{alg}}|}\right] \leq q, \tag{12}$$

where we define $0/0 = 0$. FDR is the expected ratio of false patterns in $\mathcal{E}_{\mathrm{alg}}$.

Note that an SEPM algorithm that controls FWER at significance level q also controls FDR at the same level q, but not vice versa. To guarantee that an SEPM algorithm controls FWER or FDR, we formulate this problem as multiple hypothesis testing in the following section.

4 MULTIPLE HYPOTHESIS TESTING

In this section, we first formulate an SEPM problem as multiple hypothesis testing. We then discuss the Bonferroni method and the step-up method for controlling FWER and FDR, respectively. After that, we discuss the procedure for reducing the number of hypotheses by testability, which we generalize into two-stage mining.

4.1 SEPM as Multiple Hypothesis Testing

4.1.1 Hypothesis. Section 3.3 formalized SEPM. Considering each pattern as a hypothesis, SEPM naturally fits into the framework of multiple hypothesis testing. In SEPM, a null hypothesis that corresponds to pattern e is

$$H_e^0 : \mu_e = a. \tag{13}$$

Rejecting the null hypothesis H_e^0 implies that the following alternative hypothesis is supported:

$$H_e^1 : \mu_e > a. \tag{14}$$

The alternative hypothesis H_e^1 states that pattern e is considered to be the true emerging pattern: $e \in \mathcal{E}_{\mathrm{true}}$. Whether the null hypothesis H_e^0 is rejected or not at a given significance level q is determined by the p-value of pattern e. The p-value is described in Section 4.1.2. Note that thanks to the monotonicity of the p-value, the case of $\mu_e < a$ is also covered by evaluating its null hypothesis H_e^0.

A discovered pattern $e \in \mathcal{E}_{\mathrm{alg}}$ is called a *true discovery* if e is a true emerging pattern: $e \in \mathcal{E}_{\mathrm{true}}$. On the other hand, $e \in \mathcal{E}_{\mathrm{alg}}$ is called a *false discovery*, or Type I error, if e is a false pattern: $e \in \mathcal{E}_{\mathrm{false}}$. A true emerging pattern $e \in \mathcal{E}_{\mathrm{true}}$ that fails to be rejected (not discovered) is called a Type II error. Obviously, an algorithm that rejects no hypothesis contains no false discovery

and always controls FWER and FDR at any significance level q. Such an algorithm is, however, utterly useless. In general, there is a trade-off between the number of Type I errors and the number of Type II errors. We are interested in a multiple hypothesis testing algorithm that controls FWER or FDR at q and while providing high statistical power, i.e., few Type II errors.

4.1.2 p-value. Under the null hypothesis H_e^0, the labels corresponding to pattern e follow a Bernoulli distribution with probability a. Thus, the number of positive samples given the total number of samples follows a Binomial distribution. We denote N_e and N_e^+ as the realized values of $\mathrm{Sup}(e;\mathcal{D})$ and $\mathrm{Sup}(e;\mathcal{D}^+)$. Then, the p-value of pattern e is defined as the probability that N_e^+ or more samples out of N_e samples have positively labeled under the null hypothesis H_e^0. We use p_e to denote the p-value of e, which is computed as follows:

$$\begin{aligned} p_e &= \mathbb{P}\left[\mathrm{Sup}(e;\mathcal{D}^+) \geq N_e^+ \mid \mathrm{Sup}(e;\mathcal{D}) = N_e, H_e^0\right] \\ &= \sum_{n=N_e^+}^{N_e} \binom{N_e}{n} a^n (1-a)^{N_e-n}. \end{aligned} \tag{15}$$

Obtaining the above p_e requires an exponential computation in N_e and is quite hard when N_e is large. Let $\hat{\mu}_e = N_e^+/N_e$ be the empirical positive label probability of e. Then, p_e can be approximated by the following Chernoff bound:

$$p_e^{\mathrm{C}} \leq \begin{cases} \exp(-N_e d_{\mathrm{KL}}(\hat{\mu}_e, a)) & (\hat{\mu}_e > a), \\ 1 & (\text{otherwise}), \end{cases} \tag{16}$$

where $d_{\mathrm{KL}}(p,q) = p\log(p/q) + (1-p)\log((1-p)/(1-q))$ is the KL divergence between two Bernoulli distributions with their parameters p and q. Note that the exponential factor of the Chernoff bound is optimal [8], and thus the approximation using the bound is very tight for a sufficiently large value of $|\mathcal{D}|$. In this paper, we use the above p_e^{C} as a proxy of the true p-value p_e when we need to compute p_e for $|\mathcal{D}| > 100$.

In the case of single hypothesis testing, we reject the null hypothesis H_e^0 if its p-value p_e is lower than the given significance level q; however, in the case of multiple hypothesis testing, we need to correct the significance level q based on the basis of the number of hypotheses.

4.2 Bonferroni Method for FWER

For a single null hypothesis H_e^0, the probability of a false discovery is upper-bounded by its p-value p_e. However, the probability of finding a false pattern ($e \in \mathcal{E}_{\mathrm{false}}$) with p_e smaller than any level q, converges to 1 as the number of hypotheses increases. Therefore, a multiple hypothesis testing correction is necessary for bounding FWER.

Let $e_1, \ldots, e_m$ be a set of m patterns that we would like to test. FWER can be controlled by adapting Bonferroni's correction [7], which simply divides the required significance level q by the number of patterns m. Then, the discovered patterns $\mathcal{E}_{\mathrm{alg}}$ by the Bonferroni method is

$$\mathcal{E}_{\mathrm{alg}} = \left\{ e \in \{e_1, \ldots, e_m\} : p_e \leq \frac{q}{m} \right\}, \tag{17}$$

and the Bonferroni method correctly controls FWER at significance level q.

4.3 The Step-Up Method for FDR

Unlike FWER, a simple correction factor, such as q/m, is not able to control FDR appropriately because the ratio of the expected number of false discoveries depends on all p-values of the patterns to test. For example, suppose we have tested some patterns and rejected k out of them. If all rejected patterns have low p-values, the expected number of false discoveries in the k rejected patterns is also low. In such a case, we can safely reject the next pattern even if its p-value is not very low because the average p-value of the rejected patterns is smaller than the given threshold. Conversely, if some of the rejected patterns have high p-values, we have to require a small p-value for other patterns to be rejected so as to keep the average p-value of the rejected patterns small. Consequently, in the case of controlling FDR, whether or not to reject a pattern depends on the p-values of the other patterns.

The step-up method [13] is widely used for controlling FDR. Given m patterns $e_1, \ldots, e_m$ and their p-values $p_{e_1}, \ldots, p_{e_m}$, we use $e_{(i)}$ to refer the pattern with the ith smaller p-value and $p_{(i)}$ to denote the p-value of $e_{(i)}$; namely, $p_{(1)} \leq \cdots \leq p_{(m)}$. Accordingly, the step-up method outputs the following patterns as discoveries:

$$\mathcal{E}_{\mathrm{alg}} = \left\{ e \in \{e_1, \ldots, e_m\} : p_e \leq p_{(k)} \right\}, \tag{18}$$

$$k = \operatorname*{arg\,max}_{0 \leq i \leq m} \left\{ p_{(i)} \leq \frac{q}{c(m)} \frac{i}{m} \right\}, \tag{19}$$

where $c(m)$ is an adjustment function. The correction factor in the step-up method is proportional the number of rejected hypotheses.

When each hypothesis is independent, setting $c(m) = 1$ suffices to control FDR at a significance level q, which we call the Benjamini-Hochberg (BH) method [4]. On the other hand, the Benjamini-Yekutieli (BY) method [6], which uses $c(m) = \sum_{i=1}^{m}(1/i)$, controls FDR when hypotheses have an arbitrary dependence.

4.4 Two-Stage Procedure for Improving the Statistical Power

In SEPM, each pattern is naturally associated with a hypothesis, and the number of possible patterns is exponential to the number of items. The larger the number of patterns m to test is, the weaker the power of finding significant patterns becomes, as indicated by m in the denominator of (17) and (19). It may be the case that the number of hypotheses is so large that a simple application of the Bonferroni or step-up method yields no pattern.

In the context of statistical association mining, LAMP [28] controls FWER and uses *Tarone's exclusion principle* for improving its statistical power.

Tarone's Exclusion Principle [26]: a pattern e is said to be *untestable* with respect to significance level q if the lower bound of its p-value is greater than q. Untestable patterns do not increase FWER and can be ignored before running a multiple hypothesis test.

The notion of testability is generalized in the following two-stage procedure:

Selection Stage: Find $\mathcal{E}_{\mathrm{select}} \subseteq 2^I$: each $e \in \mathcal{E}_{\mathrm{select}}$ is testable.

Testing Stage: Conduct a multiple hypothesis test on the selected testable patterns $\mathcal{E}_{\mathrm{select}}$ and output $\mathcal{E}_{\mathrm{alg}} \subseteq \mathcal{E}_{\mathrm{select}}$ where $e \in \mathcal{E}_{\mathrm{alg}}$ is rejected at a corrected significance level.

Taking this into consideration, in the next section, we propose two-stage methods for controlling FWER and FDR in SEPM. We should point out that checking testability in FDR is not a trivial exercise because the correction factor of the step-up method depends not only on the number of hypotheses but also on the p-values of all patterns. To put it differently, we cannot judge whether the lower-bound of the p-value is greater than the corrected significance level or not in the selection stage. To solve this problem, we put forward a new notion called *quasi-testability* and use it to select a subset of patterns to test.

5 PROPOSED METHODS

We propose two SEPM algorithms: LAMP-EP to control FWER and QT-LAMP-EP to control FDR. Both algorithms are based on the two-stage procedure discussed above.

5.1 LAMP-EP for FWER

LAMP-EP is a version of LAMP for statistical emerging pattern mining (SEPM). Testability is critical for reducing the number of patterns to test: under Bonferroni's correction, a pattern e is said to be untestable if the lower bound of its p-value p_e is greater than the corrected significance level q/m, where q is a given significance level and m is the number of patterns to test. In SEPM, from the definition of the p-values shown in Equation (15), the p-value of e can be lower-bounded as $p_e \geq a^{\mathrm{Sup}(e;\mathcal{D})}$, where $\mathrm{Sup}(e;\mathcal{D})$ is the number of occurrences of e in $\mathcal{D}$. Let $\psi(\tau) = a^{\tau}$. Given q and m, a pattern e is said to be *testable* with respect to q and m if $\psi(\mathrm{Sup}(e;\mathcal{D})) \leq (q/m)$. Furthermore, the untestable patterns can be safely removed from the candidates thanks to Tarone's exclusion principle [26].

In the selection stage, we find $\mathcal{E}_{\mathrm{select}} \subseteq 2^{I}$ that only consists of testable patterns. As the support of a pattern determines its testability, we want to find an appropriate threshold τ such that $\mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D})$ contains no untestable pattern, and choose $\mathcal{E}_{\mathrm{select}} = \mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D})$. We here introduce the optimal selection threshold τ^*_{FWER} defined as follows:

$$\tau^*_{\mathrm{FWER}} = \min_{\tau} \left\{ \tau : \psi(\tau) \leq \delta_{\mathrm{FWER}}(\tau; q, \mathcal{D}) \right\}, \tag{20}$$

$$\delta_{\mathrm{FWER}}(\tau; q, \mathcal{D}) = \frac{q}{|\mathcal{E}_{\mathrm{FP}}(\tau;\mathcal{D})|}, \tag{21}$$

where $\delta_{\mathrm{FWER}}(\tau; q, \mathcal{D})$ is the corrected significance level. Choosing $\mathcal{E}_{\mathrm{select}} = \mathcal{E}_{\mathrm{FP}}\left(\tau^*_{\mathrm{FWER}};\mathcal{D}\right)$ is reasonable because it is the largest set of frequent patterns that only consists of testable patterns. The following proposition clarifies the optimization problem of τ^*_{FWER}.

Proposition 5.1. *Given a dataset $\mathcal{D}$ and significance level q such that $a^{|\mathcal{D}|} < q$, there exists a unique τ^*_{FWER} such that*

$$\psi(\tau^*_{\mathrm{FWER}} - 1) > \delta_{\mathrm{FWER}}\left(\tau^*_{\mathrm{FWER}} - 1; q, \mathcal{D}\right), \tag{22}$$

$$\psi(\tau^*_{\mathrm{FWER}}) \leq \delta_{\mathrm{FWER}}\left(\tau^*_{\mathrm{FWER}}; q, \mathcal{D}\right), \tag{23}$$

Proof. Let $\mathrm{LHS}(\tau) = \phi(\tau)$ and $\mathrm{RHS}(\tau) = \delta_{\mathrm{FWER}}(\tau; q, \mathcal{D})$. $\mathrm{LHS}(\tau)$ and $\mathrm{RHS}(\tau)$ are decreasing and increasing functions of τ, respectively. Namely, if τ satisfies $\mathrm{LHS}(\tau) \leq \mathrm{RHS}(\tau)$, all $\tau' > \tau$ also satisfy the same inequation. For $\tau = 0$, $\mathrm{LHS}(\tau) > \mathrm{RHS}(\tau)$ holds because $\psi(0) = 1$ and $\delta_{\mathrm{FWER}}(0; q, \mathcal{D}) \leq q < 1$. For $\tau = |\mathcal{D}|$, $\mathrm{LHS}(\tau) < \mathrm{RHS}(\tau)$ holds because $\psi(|\mathcal{D}|) = a^{|\mathcal{D}|} < q = \delta_{\mathrm{FWER}}(|\mathcal{D}|; q, \mathcal{D})$, where we have used the fact that $|\mathcal{E}_{\mathrm{FP}}(|\mathcal{D}|;\mathcal{D})| = |\{\phi\}| = 1$ in the last transformation. Combining the above facts, we see that there exists the unique threshold value τ such that $\mathrm{LHS}(\tau) > \mathrm{RHS}(\tau)$ and $\mathrm{LHS}(\tau + 1) \leq \mathrm{RHS}(\tau + 1)$. □

It is easy to see that τ^*_{FWER} in Eq. (22) and Eq. (23) corresponds the selection threshold defined in Eq. (20). Due to the monotonicity of each side of the inequality, τ^*_{FWER} can be found by using a bisection search.

Taking the above fact into consideration, we propose LAMP-EP as the following two-stage procedure:

Selection Stage: Find τ^*_{FWER} and obtain $\mathcal{E}_{\mathrm{select}} = \mathcal{E}_{\mathrm{FP}}\left(\tau^*_{\mathrm{FWER}};\mathcal{D}\right)$ by using a frequent pattern mining algorithm.

Testing Stage: Conduct a multiple test with the Bonferroni's method for $\mathcal{E}_{\mathrm{select}}$ and output the rejected patterns as $\mathcal{E}_{\mathrm{alg}}$.

The LAMP-EP correctly controls FWER at a significance level q.

5.2 QT-LAMP-EP for FDR

In the case of FWER, thanks to Tarone's exclusion principle, we are able to remove untestable patterns from the candidate patterns and conduct Bonferroni's correction on the remaining patterns. Gilbert [11] showed that a Tarone-like exclusion principle can also be used for controlling FDR when the patterns are independent. However, when the patterns are dependent, which is the case in SEPM, it is not known whether the same principle applies or not. In order to improve the statistical power under a controlled FDR, we need a new exclusion principle that reduces the number of patterns before conducting the step-up correction. In the following, we propose an *unbiasedness condition* that is sufficient for controlling FDR. After that, we propose *quasi-testability* for optimizing statistical power. Following the introduction of these notions, we describe QT-LAMP-EP that controls FDR.

Let $\mathcal{E}_{\mathrm{select}}$ be a set of selected patterns in the selection stage. $\mathcal{E}_{\mathrm{select}}$ is said to be *unbiased* with respect to dataset $\mathcal{D}'$ if, for any transaction $(x, y) \in \mathcal{D}'$,

$$\mathbb{P}[y \mid e \subseteq x] = \mathbb{P}[y \mid e \subseteq x, \mathcal{E}_{\mathrm{select}}]. \tag{24}$$

If an unbiased $\mathcal{E}_{\mathrm{select}}$ can be obtained in the selecting stage, we can apply the step-up method for $\mathcal{E}_{\mathrm{select}}$ in the testing stage because p-value conditioned on the selection is correctly controlled: the resulting two-stage procedure always controls FDR.

To guarantee unbiasedness, we first split the given dataset $\mathcal{D}$ into two sub-datasets: the *calibration dataset* $\mathcal{D}_{\mathrm{carib}}$ and the *main datasets* $\mathcal{D}_{\mathrm{main}}$. In the selection stage, we first compute a selection threshold τ_{FDR} from $\mathcal{D}_{\mathrm{carib}}$ and set $\mathcal{E}_{\mathrm{select}} = \mathcal{E}_{\mathrm{FP}}(\tau_{\mathrm{FDR}};\mathcal{D}_{\mathrm{main}})$. Then, we conduct the step-up method for $\mathcal{E}_{\mathrm{select}}$ using $\mathcal{D}_{\mathrm{main}}$ in the testing stage. This two-stage procedure satisfies the above unbiasedness condition with respect to $\mathcal{D}_{\mathrm{main}}$; because of the assumed i.i.d. property of the transactions in $\mathcal{D}$, τ_{FDR} is determined independent of the realization of the main dataset $\mathcal{D}_{\mathrm{main}}$. Therefore, whether or not a pattern e is selected is determined solely on its support size. Since the p-value is defined by the conditional probability on $\mathrm{Sup}(e;\mathcal{D}) = N_e$ (Section 4.1.2) and no information on the labels of these pattern is exploited, the p-value correctly represents the tail probability if e is a null hypothesis.

Inspired by testability, we propose a new notion called *quasi-testability*: e is *quasi-testable* with respect to q, $\hat{k}$, and $\hat{m}$ if

$$\psi(\mathrm{Sup}(e; \mathcal{D}_{\mathrm{carib}})) \leq \frac{q}{c(\hat{m})} \frac{\hat{k}}{\hat{m}}. \tag{25}$$

If we can set $\hat{k} = k$ and $\hat{m} = m$ in Equation (19), quasi-testability is a sufficient condition for the pattern not to be rejected by the step-up method. We estimate these values from $\mathcal{D}_{\mathrm{carib}}$. In the selection stage, we find $\mathcal{E}_{\mathrm{select}} \subseteq 2^I$ such that $\mathcal{E}_{\mathrm{select}}$ consists of only quasi-testable patterns. Similar to the LAMP-EP (Section 5.1) that selects the largest set of testable patterns, we would like to find τ^*_{FDR} defined as follows:

$$\tau^*_{\mathrm{FDR}} = \min_{\tau} \{\tau : \psi(\tau) \leq \delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}})\}, \tag{26}$$

$$\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}}) = \frac{q}{c(|\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{carib}})|)} \frac{\hat{k}(\tau; \mathcal{D}_{\mathrm{carib}})}{|\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{carib}})|}, \tag{27}$$

where $\hat{k}(\tau; \mathcal{D}_{\mathrm{carib}})$ is an estimator of k of Equation (19), which is computed by conducting the step-up correction on $\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{carib}})$ by using $\mathcal{D}_{\mathrm{carib}}$, and $\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}})$ is the corrected significance level. The following proposition clarifies the optimization of τ^*_{FDR}:

Proposition 5.2. *Given a dataset $\mathcal{D}_{\mathrm{carib}}$ and a significance level q such that $a^{|\mathcal{D}_{\mathrm{carib}}|} < q$, there exists a threshold τ_{FDR} such that*

$$\psi(\tau_{\mathrm{FDR}} - 1) > \delta_{\mathrm{FDR}}(\tau_{\mathrm{FDR}} - 1; q, \mathcal{D}_{\mathrm{carib}}) \tag{28}$$

$$\psi(\tau_{\mathrm{FDR}}) \leq \delta_{\mathrm{FDR}}(\tau_{\mathrm{FDR}}; q, \mathcal{D}_{\mathrm{carib}}) \tag{29}$$

and τ^*_{FDR} defined by Eq. (26) is the minimum of such τ_{FDR}.

Proof. Let $\mathrm{LHS}(\tau) = \phi(\tau)$ and $\mathrm{RHS}(\tau) = \delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}})$. Because $\psi(0) = 1$ and $\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}}) \leq q < 1$ hold, $\mathrm{LHS}(0) > \mathrm{RHS}(0)$ holds. For $\tau = |\mathcal{D}_{\mathrm{carib}}|$, $\mathrm{LHS}(\tau) < \mathrm{RHS}(\tau)$ holds because $\psi(|\mathcal{D}_{\mathrm{carib}}|) = a^{|\mathcal{D}_{\mathrm{carib}}|}$ and $\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}}) = q$. From the above facts, there exists at least one τ such that $\mathrm{LHS}(\tau) > \mathrm{RHS}(\tau)$ and $\mathrm{LHS}(\tau + 1) \leq \mathrm{RHS}(\tau + 1)$. By definition, τ^*_{FDR} is the minimum of such τ_{FDR}. □

Note that the above τ_{FDR} is not necessary unique. This uniqueness is due to the inherent nature of FDR: the corrected significance level $\delta_{\mathrm{FDR}}(\tau; q, \mathcal{D}_{\mathrm{carib}})$ is not monotone to the addition of patterns. This non-monotonicity is closely related to the following fact. In the step-up method, non-significant patterns can become significant if we add more patterns with very small p-values. Unfortunately, finding τ^*_{FDR}, which results a largest set of quasi-testable patterns among the possibly non-unique values of τ_{FDR}, requires an incremental search over τ [21, 25], which is not very efficient when no early termination is applied. In practice, we observed that τ_{FDR} is usually unique and thus corresponds to τ^*_{FDR} (see Figure 3 in Section 6). Here, therefore, we will look for one of τ_{FDR} instead of the τ^*_{FDR} and use the found τ_{FDR} to select the patterns in the selection stage. Note that a bisection search over threshold values can be used for finding one of τ_{FDR}.

Finally, we propose QT-LAMP-EP as the following two-stage procedure:

Selection Stage: Find τ_{FDR} by using calibration dataset $\mathcal{D}_{\mathrm{carib}}$ and obtain $\mathcal{E}_{\mathrm{select}} = \mathcal{E}_{\mathrm{FP}}(\tau_{\mathrm{FDR}}; \mathcal{D}_{\mathrm{main}})$ by using a frequent pattern mining algorithm.

Testing Stage: Conduct a multiple hypothesis test with the step-up method for $\mathcal{E}_{\mathrm{select}}$ with dataset $\mathcal{D}_{\mathrm{main}}$ and output the rejected patterns as $\mathcal{E}_{\mathrm{alg}}$.

The following Theorem 5.3 states that QT-LAMP-EP strictly controls FDR for a given error rate q.

Theorem 5.3. *Assume that $\mathcal{E}_{\mathrm{select}}$ satisfies the unbiased property (Ineq. (24)). With $c(m) = \sum_{i=1}^{m}(1/i)$, The FDR of QT-LAMP-EP is q or smaller.*

Proof. Let the total patterns be 2^I. Let $S \subseteq 2^I$ be a subset of patterns. Let $Q = V/R$ be the rate of false discoveries, where V and R are the numbers of rejected true null hypotheses and rejected null hypotheses, respectively. To prove Theorem 5.3, we use the following lemma.

Lemma 5.4. (Theorem 1.3 in Benjamini and Yekutieli [6]) *For arbitrary $\delta \in [0, 1]$, if $\mathbb{P}[p_e \leq \delta \mid \mathcal{E}_{\mathrm{select}} = S] \leq \delta$ holds for a true hypothesis, then $\mathbb{E}[Q \mid \mathcal{E}_{\mathrm{select}} = S] \leq q$ holds by applying the step-up method with the BY correction at the second stage.*

Note that $\mathbb{P}[p_e \leq q \mid \mathcal{E}_{\mathrm{select}} = S] \leq q$ holds if $\mathcal{E}_{\mathrm{select}}$ has the unbiased property. Therefore, the FDR is bounded as follow:

$$\begin{aligned}
\mathrm{FDR} = \mathbb{E}[Q] &= \sum_{S \subseteq 2^I} \mathbb{E}[Q \mid \mathcal{E}_{\mathrm{select}} = S]\mathbb{P}[\mathcal{E}_{\mathrm{select}} = S] \\
&\leq \sum_{S \subseteq 2^I} q\mathbb{P}[\mathcal{E}_{\mathrm{select}} = S] \qquad \text{(By Lemma 5.4)} \\
&= q.
\end{aligned}$$

□

6 EXPERIMENTS

This section describes the results of computer simulations. First, we show the results for a synthetic dataset on which we measured the false discovery rate. Second, we show the results for eight real-world datasets to verify the statistical power of our algorithms.

6.1 Hardware and Software

We used a Linux machine with two 12-Core 2.40GHz Intel Xeon (E5-2620 v3) CPUs and 132GB of memory for running all the experiments. LAMP-EP and QT-LAMP-EP each used an FPM algorithm as a building block. We implemented our algorithms on top of LCM++[1], an open source C++ implementation of the LCM [29] algorithm. LCM is an award-winning[2] algorithm for frequent pattern mining and is known as one of the fastest one for FPM. LCM++ is a little bit slower than the original implementation of LCM but has improved readability; it enables each of our experiment that involves from thousands to a half a million transactions to be completed within a day. For the ease of computation, we restricted our interest to closed itemsets that had no superset of the same occurrence. We used a bisection search over threshold values to find τ^*_{FWER} and τ_{FDR} for LAMP-EP and QT-LAMP-EP, respectively.

[1] https://code.google.com/archive/p/lcmplusplus/
[2] http://fimi.ua.ac.be/fimi04/

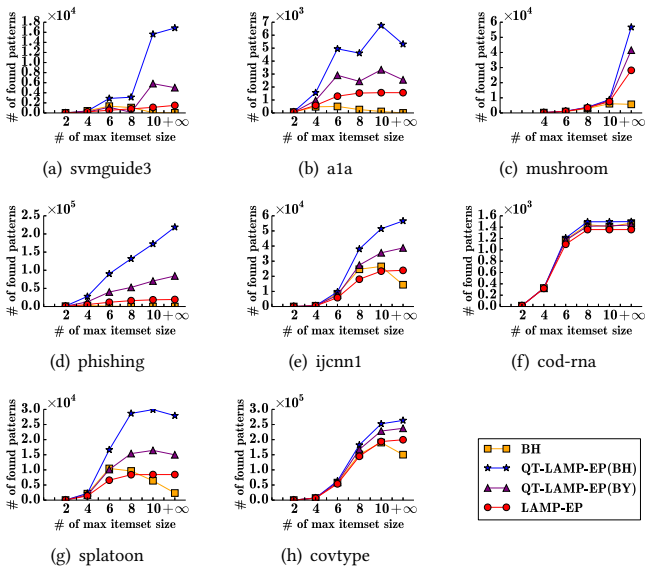


Figure 1: Number of discovered patterns. The horizontal axis indicates the maximum size of the patterns to search, where $+\infty$ indicates a limitless procedure where all patterns are searched.

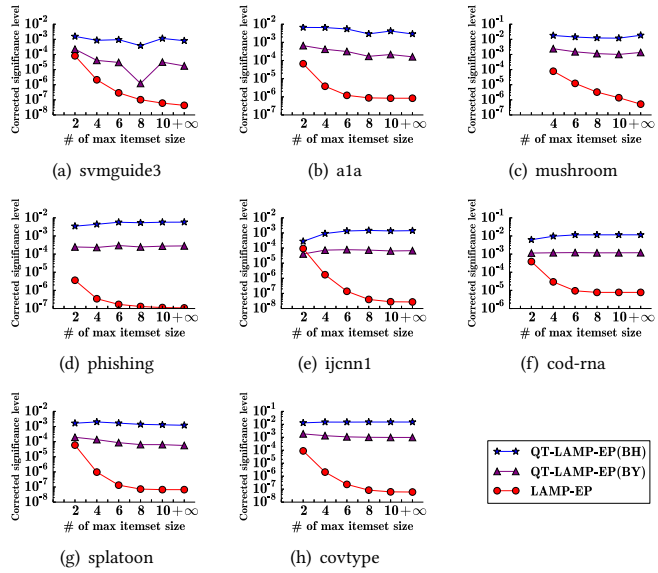


Figure 2: Corrected significance level of algorithms. The horizontal axis is the same as in Figure 1, and the vertical axis is the significance level. The LAMP-EP results show the corrected significance level. The QT-LAMP-EP results show $qk/(c(m)m)$, where k is the number of rejected hypotheses among m ones. We omit presenting the significance of BH because its significance level was too low to draw in the figure.

6.2 Experimental Settings

Throughout the experiments, the significance level q was set to be 0.05 for both FWER and FDR mining.

We compared the following algorithms as to their numbers of discoveries and corrected significance levels: **EPM** is a traditional emerging pattern mining algorithm for finding $\mathcal{E}_{\mathrm{EP}}(a; \mathcal{D})$; we set its minimum support to 10 for the sake of a short enough computation. **LAMP-EP** and **QT-LAMP-EP** are our methods. We tested the BH and BY corrections (i.e., $c(m) = 1$ and $\sum_{i=1}^{m}(1/i)$, resp.) for QT-LAMP-EP. **BH** is the standard BH correction that where the number of hypotheses m is all the possible patterns 2^I. We did not use the standard Bonferroni's method that tests all the patterns because it always finds the same number or fewer discoveries than BH does. In QT-LAMP-EP, we used 80% of $\mathcal{D}$ as the main dataset $\mathcal{D}_{\mathrm{main}}$. The calibration dataset $\mathcal{D}_{\mathrm{carib}}$ was resampled from the other 20% of $\mathcal{D}$ until its size reaches the size of $\mathcal{D}_{\mathrm{main}}$. Namely, we spend 20% of the dataset to determine an appropriate value of τ_{FDR}. Although this splitting generated a randomness in the algorithm, the randomness had little effect on the large databases. The other algorithms used the entire $\mathcal{D}$.

6.3 Synthetic Dataset

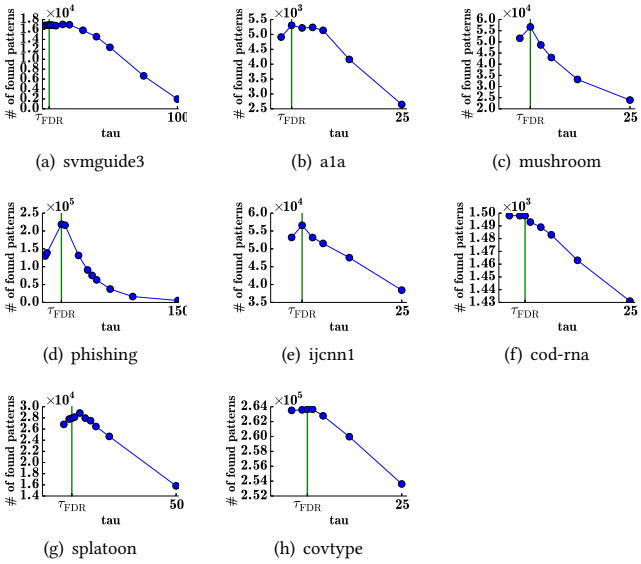
To show how many true and false discoveries were found by the algorithms listed in Section 6.2, we conducted simulations with a synthetic dataset. The dataset consisted of 100,000 transactions that involved itemsets with $|I| = 100$. 10% of the transactions contained some of items $\{1, \ldots, 10\}$ and $\mathbb{P}[y \mid x] = 0.7$ for these transactions. The other 90% of the transactions contained items $\{11, 12, \ldots, 100\}$ and $\mathbb{P}[y \mid x] = 0.5$. The value of a was set to be 0.5, and thus, the patterns that contained items in $\{11, 12, \ldots, 100\}$ were false patterns.

Table 2 shows the results of the experiment. EPM, which lacks a statistical assessment, finds the largest set of true discoveries but suffers a very high family-wise error rate. LAMP-EP controls both FDR and FWER. QT-LAMP-EP with both BH and BY adjustments factor maintains FDR at a level less than q.

Necessity of BY adjustment: The BY adjustment (i.e., $c(m) = \sum_{i=1}^{m} 1/i$) is theoretically required for controlling FDR. However, in Benjamini and Yekutieli [6], it is noted that the BY adjustment is "very often unneeded, and yields too conservative a procedure". This remark is consistent with our result that QT-LAMP-EP (BY) has much smaller FDR than q: even on this synthetic dataset where the correlation among patterns is very large, the BH adjustment (i.e., $c(m) = 1$) alone is enough to control FDR.

Table 2: Performance of the procedures. TDs and FDs correspond to true and false discoveries, respectively. The results are empirical averages over 100 runs.

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



(a) svmguide3 (b) a1a (c) mushroom

(d) phishing (e) ijcnn1 (f) cod-rna

(g) splatoon (h) covtype

Figure 3: Appropriateness of τ_{FDR}. The horizontal axis is the minimum support τ of FPM, and the vertical axis is the number of discovered patterns when we conducted the BH method on the frequent patterns $\mathcal{E}_{\mathrm{FP}}(\tau; \mathcal{D}_{\mathrm{main}})$. The horizontal line indicates τ_{FDR} that is optimized by QT-LAMP-EP.

Gap between the actual FDR and q: There are several reasons for the gap between the actual FDR and $q = 0.05$: First, the step-up method actually controls FDR at a level $q(m_0/m) \leq q$, where m_0 is the number of true null hypotheses (i.e., insignificant patterns) that are selected in the first stage [5]. Second, the discreteness of the Binomial probability implies that $\mathbb{P}[p_e \leq \delta]$ is smaller than δ. Third, since the p-values in our experiment were highly dependent, the probability of having an extremely low p-value was not very large. For these reasons, it is not surprising that the actual false discovery rate of QT-LAMP-EP was smaller than q.

6.4 Real-world Datasets

The simulations involved eight datasets. Mushroom dataset was obtained from the UCI repository[3]. The Splatoon[4] dataset consisted of the results of online multi-player games. We gathered the results of about 400,000 Splatoon matches on Stat.ink[5] from October 31, 2015 to January 30, 2016. We converted the players' weapons, ranks, and the features related to the battle arena into items. The other six datasets (A1a, Cod-RNA, Covtype, IJCNN1, Phishing, and SVMGuide3) were obtained from the libSVM dataset repository[6]. The real-valued features of some datasets were divided into two classes by thresholding at the median, and each class was converted into an item. The statistics of the datasets are shown in Table 3. We set the value of a to 0.3 for A1a, IJCNN1, and SVMGuide3, 0.4 for Cod-RNA, 0.5 for Mushroom, 0.55 for Covtype, 0.6 for Splatoon, and 0.8 for Phishing. These values are set so as the number of the found patterns are modest.

Figure 1 shows the number of found patterns. The BH procedure that takes all possible patterns into consideration performed worst. This means our algorithms were effective at boosting the statistical power of finding patterns. For all datasets, when the maximum size of patterns to search was large, QT-LAMP-EP with both the BH and BY correction outperformed LAMP-EP. QT-LAMP-EP with the BH adjustment always finds more patterns than the QT-LAMP-EP with the BY adjustment; this result is natural since the BH adjustment admits a larger significance level. The advantage of QT-LAMP-EP is even larger in terms of the significance level (Figure 2). In the limitless case, QT-LAMP-EP (BY), which strictly controls FDR (Theorem 5.3), admitted about a 10^2 to 10^3 times larger significance level than that of LAMP-EP. QT-LAMP-EP (BH), which is very likely to control FDR as shown in the synthetic data, even admitted a 10 times larger significance level on many datasets.

6.5 Appropriateness of the Selection Threshold

To verify the appropriateness of the selection threshold τ_{FDR} in terms of its power to make discoveries, we varied the minimum support τ and conducted a BH over all frequent patterns with τ. Figure 3 shows the number of found patterns as a function of τ: Here, the τ_{FDR} selected by QT-LAMP-EP does not always maximize the number of patterns but was always a very close-to-optimal choice of τ. This result empirically supports the proposed criteria based on quasi-testability.

6.6 Sensitivity on the Value of a

We also conducted simulations with different values of the positive label probability threshold a and no maximum pattern sizes. Figure 4 shows the results of that experiment. The definition of SEPM means that it natural that the number of found patterns is the decreasing as a increases. The fact that QT-LAMP-EP finds more pattern than LAMP-EP is consistent with all the values of a and datasets.

7 CONCLUSION

We studied the problem of finding emerging patterns that are significant in the sense of multiple testing. We devised procedures, called LAMP-EP and QT-LAMP-EP to control FWER and FDR. The

Table 3: Statistics of the datasets. $|\mathcal{D}|$ and $|I|$ are the numbers of transactions and items (features), respectively. Avg $|x|$ is the averaged number of items in a datapoint.

dataset	$\lvert\mathcal{D}\rvert$	$\lvert I\rvert$	Avg $\lvert x\rvert$	$\lvert\mathcal{D}^+\rvert/\lvert\mathcal{D}\rvert$
svmguide3	1,243	42	20.9	0.24
a1a	1,605	174	13.9	0.25
mushroom	8,124	117	22.0	0.48
phishing	11,055	813	30.0	0.56
ijcnn1	49,990	44	13.0	0.10
codrna	59,535	16	8.0	0.33
splatoon	404,515	54	11.0	0.50
covtype	581,012	64	11.9	0.49

[3]https://archive.ics.uci.edu/ml/datasets/Mushroom
[4]https://www.nintendo.com/games/detail/splatoon-wii-u
[5]https://stat.ink/
[6]https://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/

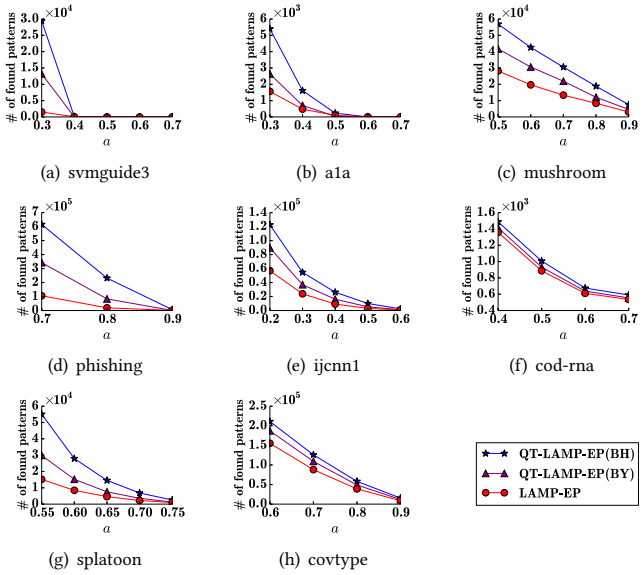


Figure 4: Number of patterns found as a function of a. The horizontal axis is the value of a, and the vertical axis is the corresponding number of found patterns.

difficulty of multiple testing stems from the fact that the number of hypotheses is exponential to the size of the itemset; this motivated us to select a subset of patterns to test. In controlling FWER, as is done in previous papers [25, 28], our procedure selects testable hypotheses. For controlling FDR, we proposed the criteria of quasi-testability that effectively eliminates most of the hypotheses by estimating the corrected significance level. Notably, ours is the first method that controls FDR in combinatorial pattern mining with adaptive selection of hypotheses. It is not very difficult to apply our QT-LAMP-EP to other pattern mining tasks such as statistical association mining. The following are important directions of future works:

A cleverer data splitting: in deriving τ_{FDR}, we used a calibration dataset that is different from the main dataset. This restriction comes from the requirement of unbiasedness (Ineq. (24)). If one can fully utilize the entire dataset for the testing by relaxing the requirement, it will increase the efficiency of how the dataset is used.

A more efficient computation of τ_{FDR}: Sugiyama et al. [25] discussed that an incremental search with an efficient early termination rule runs faster than a bisection search. An incremental search algorithm for QT-LAMP-EP, preferably with a one-pass modification [21], would enable QT-LAMP-EP to be applied to datasets with billions of transactions. Alternatively, a bisection search that starts with an estimated lower bound [30] of τ_{FDR} can be much faster since it can avoid computing FPs with low support thresholds.

8 ACKNOWLEDGEMENT

This work was supported in part by JSPS KAKENHI Grant number 17K12736 and 15H05711. This work was supported in part by the Research and Development on Real World Big Data Integration and Analysis Program of MEXT. We thank Masashi Toyoda for letting us use their computing resources. We thank anonymous reviewers for their insightful comments.

REFERENCES

[1] Charu C. Aggarwal and Jiawei Han (Eds.). 2014. *Frequent Pattern Mining.* Springer.
[2] James Bailey, Thomas Manoukian, and Kotagiri Ramamohanarao. 2002. *Fast Algorithms for Mining Emerging Patterns.* Springer Berlin Heidelberg, Berlin, Heidelberg, 39–50.
[3] Roberto J. Bayardo, Jr. 1998. Efficiently Mining Long Patterns from Databases. In *Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data (SIGMOD '98).* ACM, New York, NY, USA, 85–93.
[4] Yoav Benjamini and Yosef Hochberg. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 57, 1 (1995), 289–300.
[5] Yoav Benjamini, Abba M. Krieger, and Daniel Yekutieli. 2006. Adaptive linear step-up procedures that control the false discovery rate. *Biometrika* 93, 3 (2006), 491–507.
[6] Yoav Benjamini and Daniel Yekutieli. 2001. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.* 29, 4 (08 2001), 1165–1188.
[7] C. E. Bonferroni. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* 8 (1936), 3–62.
[8] Amir Dembo and Ofer Zeitouni. 1998. *Large deviations techniques and applications.* Springer, New York, Berlin, Heidelberg.
[9] Guozhu Dong and James Bailey. 2012. *Contrast Data Mining: Concepts, Algorithms, and Applications* (1st ed.). Chapman & Hall/CRC.
[10] Guozhu Dong and Jinyan Li. 1999. Efficient Mining of Emerging Patterns: Discovering Trends and Differences. In *Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 15-18, 1999.* 43–52.
[11] Peter B. Gilbert. 2005. A Modified False Discovery Rate Multiple-Comparisons Procedure for Discrete Data, Applied to Human Immunodeficiency Virus Genetics. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 54, 1 (2005), 143–158.
[12] Sami Hanhijärvi. 2011. *Multiple Hypothesis Testing in Pattern Discovery.* Springer Berlin Heidelberg, Berlin, Heidelberg, 122–134.
[13] Yosef Hochberg. 1988. A sharper Bonferroni procedure for multiple tests of significance. *Biometrika* 75, 4 (1988), 800.
[14] Adam Kirsch, Michael Mitzenmacher, Andrea Pietracaprina, Geppino Pucci, Eli Upfal, and Fabio Vandin. 2012. An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets. *J. ACM* 59, 3 (2012), 12:1–12:22.
[15] Stéphane Lallich, Olivier Teytaud, and Elie Prudhomme. 2006. *Statistical inference and data mining: false discoveries control.* Physica-Verlag HD, Heidelberg, 325–336.
[16] Stephane Lallich, Olivier Teytaud, and Elie Prudhomme. 2007. *Association Rule Interestingness: Measure and Statistical Validation.* Springer Berlin Heidelberg, Berlin, Heidelberg, 251–275.
[17] Felipe Llinares-López, Mahito Sugiyama, Laetitia Papaxanthos, and Karsten M. Borgwardt. 2015. Fast and Memory-Efficient Significant Pattern Mining via Permutation Testing. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Sydney, NSW, Australia, August 10-13, 2015.* 725–734.
[18] Cécile Low-Kam, Chedy Raïssi, Mehdi Kaytoue, and Jian Pei. 2013. Mining Statistically Significant Sequential Patterns. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013.* 488–497.
[19] Shihong Mao and Guozhu Dong. 2005. Discovery of Highly Differentiative Gene Groups from Microarray Gene Expression Data Using the Gene Club Approach. *Journal of Bioinformatics and Computational Biology* 03, 06 (2005), 1263–1280.
[20] Nicolai Meinshausen, Marloes H. Maathuis, and Peter Bhlmann. 2011. Asymptotic optimality of the Westfall–Young permutation procedure for multiple testing under dependence. *Ann. Statist.* 39, 6 (12 2011), 3369–3391.
[21] Shin-ichi Minato, Takeaki Uno, Koji Tsuda, Aika Terada, and Jun Sese. 2014. *A Fast Method of Statistical Assessment for Combinatorial Hypotheses Based on Frequent Itemset Enumeration.* Springer Berlin Heidelberg, Berlin, Heidelberg, 422–436.
[22] Petra Kralj Novak, Nada Lavrač, and Geoffrey I. Webb. 2009. Supervised Descriptive Rule Discovery: A Unifying Survey of Contrast Set, Emerging Pattern and Subgroup Mining. *J. Mach. Learn. Res.* 10 (June 2009), 377–403.
[23] Laetitia Papaxanthos, Felipe Llinares-López, Dean A. Bodenham, and Karsten M. Borgwardt. 2016. Finding significant combinations of features in the presence of categorical covariates. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain.* 2271–2279.

[24] Matteo Riondato and Fabio Vandin. 2014. Finding the True Frequent Itemsets. In *Proceedings of the 2014 SIAM International Conference on Data Mining, Philadelphia, Pennsylvania, USA, April 24-26, 2014*. 497–505.

[25] Mahito Sugiyama, Felipe Llinares-López, Niklas Kasenburg, and Karsten M. Borgwardt. 2015. Significant Subgraph Mining with Multiple Testing Correction. In *Proceedings of the 2015 SIAM International Conference on Data Mining, Vancouver, BC, Canada, April 30 - May 2, 2015*. 37–45.

[26] R. E. Tarone. 1990. A Modified Bonferroni Method for Discrete Data. *Biometrics* 46, 2 (1990), 515–522.

[27] Aika Terada, Hanyoung Kim, and Jun Sese. 2015. High-speed Westfall-young Permutation Procedure for Genome-wide Association Studies. In *Proceedings of the 6th ACM Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '15)*. ACM, New York, NY, USA, 17–26.

[28] Aika Terada, Mariko Okada-Hatakeyama, Koji Tsuda, and Jun Sese. 2013. Statistical significance of combinatorial regulations. *Proceedings of the National Academy of Sciences* 110, 32 (2013), 12996–13001.

[29] Takeaki Uno, Tatsuya Asai, Yuzo Uchida, and Hiroki Arimura. 2003. LCM: An Efficient Algorithm for Enumerating Frequent Closed Item Sets. In *FIMI '03, Frequent Itemset Mining Implementations, Proceedings of the ICDM 2003 Workshop on Frequent Itemset Mining Implementations, 19 December 2003, Melbourne, Florida, USA*.

[30] Matthijs van Leeuwen and Antti Ukkonen. 2014. Fast Estimation of the Pattern Frequency Spectrum. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2014, Nancy, France, September 15-19, 2014. Proceedings, Part II*. 114–129. DOI : http://dx.doi.org/10.1007/978-3-662-44851-9_8

[31] Geoffrey I. Webb. 2007. Discovering Significant Patterns. *Machine Learning* 68, 1 (2007), 1–33.

[32] P.H. Westfall and S.S. Young. 1993. *Resampling-Based Multiple Testing: Examples and Methods for P-Value Adjustment*. Wiley.