A Rigorous Statistical Approach for Identifying Significant Itemsets

Adam Kirsch Michael Mitzenmacher Andrea Pietracaprina Geppino Pucci
Eli Upfal Fabio Vandin

ABSTRACT

As advances in technology allow for the collection, storage, and mining of vast amounts of data, the task of screening and assessing the significance of the discovered patterns is becoming a major challenge in data mining applications. In this work, we address significance in the context of frequent itemset mining. Specifically, we develop a novel methodology to identify a meaningful support threshold s for a dataset, such that the family of frequent itemsets with respect to s embodies a substantial deviation from what would be expected in a random dataset, hence these itemsets can be flagged as significant. Our methodology hinges on a Poisson approximation of the distribution of the number of frequent itemsets of a given size, which is the main theoretical result of the paper. A crucial feature of our approach is that, unlike previous work, it takes into account the entire dataset rather than individual discoveries, hence it is able to distinguishing between significant observations and random fluctuations in data, thus resulting in fewer false discoveries. Extensive experiments are reported that substantiate the effectiveness of our methodology.

1. Introduction

The discovery of frequent itemsets in transactional datasets is regarded as a fundamental primitive that arises in the mining of association rules and in many other scenarios [14, 22]. In its original formulation, the problem requires that given a dataset $\mathcal{D}$ of transactions over a set of items $\mathcal{I}$, and a support threshold s, all itemsets $X \subseteq \mathcal{I}$ with support at least s (i.e., contained in at least s transactions) be discovered and returned in output. These high-support itemsets are referred to as *frequent itemsets*.

Since the pioneering paper by Agrawal et al. [2], a vast literature has flourished proposing variants of the problem, studying foundational issues, and presenting novel algorithmic strategies or clever implementations of known strategies (see, e.g., [10, 11]), but many problems remain open [13]. In particular, assessing the significance of the discovered itemsets, or equivalently, flagging statistically significant discoveries with a limited number of false positive outcomes, is still poorly understood and remains one of the most challenging problems in this area.

The classical framework requires the user to decide what is significant by specifying the support threshold s. Unless specific domain knowledge is available, the choice of such a threshold is often arbitrary and unlikely to guarantee that the frequent itemsets discovered be significant [14,22]. While an aggressive high threshold may disregard significant itemsets (false negatives), a conservative low threshold may lead to a large number of spurious discoveries (false positives) that would undermine the success of subsequent analysis.

In this paper, we tackle this problem by introducing a suitable notion of significance, which lends itself to a more rigorous treatment. Specifically, we will flag as significant a population of itemsets extracted with respect to a certain threshold if some global characteristics of the population deviate considerably from what would be expected if the dataset were generated at random.

1.1 Related Work

A number of works have explored various notions of significant itemsets and approaches to their discovery. Below, we review those most relevant to our approach and refer the reader to [13, Section 3] for further references. The paper [1] relates the significance of an itemset X to the quantity $((1-v(X))/(1-\mathbf{E}[v(X)]))\cdot(\mathbf{E}[v(X)]/v(X))$, where $v(X)$ represents the fraction of transactions containing some but not all of the items of X, and $\mathbf{E}[v(X)]$ represents the expectation of $v(X)$ in a random dataset where items occur in transactions independently. This ratio provides an empirical measure of the correlation among the items of X that, according to [1], is more effective than absolute support. In [7, 8, 21], the significance of an itemset is measured as the ratio R between its actual support and its expected support in a random dataset. In order to make this measure more accurate for small supports, [7, 8] proposes smoothing the ratio R using an empirical Bayesian approach. Bayesian analysis

is also employed in [19] to derive subjective measures of significance of patterns (e.g., itemsets) based on how strongly they "shake" a system of established beliefs.

A statistical approach for identifying significant itemsets is presented in [20], where the measure of interest for an itemset is defined as the degree of dependence among its constituent items, which is assessed through a χ^2 test. Unfortunately, as reported in [7, 8], there are technical flaws in the applications of the statistical test in [20]. Nevertheless, [20] pioneered the quest for a rigorous framework for addressing the discovery of significant itemsets.

A common drawback of the aforementioned works is that they assess the significance of each itemset *in isolation*, rather than taking into account the *global* characteristics of the dataset from which they are extracted. In fact, if the number of itemsets considered by the analysis is large, even in a purely random dataset some of them are likely to be flagged as significant if considered in isolation. This is a well known phenomenon arising when performing multiple hypothesis testing [5, 16], and is illustrated by the following example.

Consider a simple dataset of 50M transactions over 10,000 items, where each transaction contains exactly 2 items. If we assume that the pair of items occurring in any given transaction is chosen uniformly at random from the $\approx$ 50M pairs, the expected number of occurrences of each pair is 1, but simple combinatorics shows that with probability close to 1, there will be at least one pair that appears in at least 6 transactions. Any sound strategy for mining significant itemsets should not highlight this pair as significant since the deviation from its expectation is just a random fluctuation. On the other hand, suppose we modify the collection of transactions by picking another 5 pairs of items and adding 6 transactions for each of the pairs, for a total of 30 new transactions. Now, the probability of having a total of 6 pairs with support at least 6 in (roughly) 50 million randomly generated transactions is less than $1/100$, so we argue that a good strategy for discovering significant itemsets should flag the occurrence of these 6 pairs as a statistically surprising event. However, the absolute support and the dependency test value of [20] for the new pairs is about the same as for the one pair with support at least 6 in the random dataset, and so these measures cannot distinguish between the two cases.

A few works attempt at accounting for the global structure of the dataset in the context of frequent itemset mining. The authors of [9] propose a Markov chain-based approach to generate a random dataset that has similar marginal distribution properties as a given real dataset (namely, identical transaction lengths and identical frequencies of the individual items[1]). The work suggests comparing the outcomes of a number of data mining tasks, frequent itemset mining among the others, in the real and the randomly generated datasets in order to establish whether the real datasets exhibit any significant global structure. However, such an assessment is carried out in a purely qualitative fashion without rigorous statistical grounding.

[1]By *frequency* of an item, we mean the fraction of transactions containing the item.

The problem of spurious discoveries when mining significant patterns is studied in [5]. The paper is concerned with the discovery of significant pairs of items, where significance is measured through the p-value, that is, the probability of occurrence of the observed support in a random dataset. Significant pairs are those whose p-values are below a certain threshold that can be suitably chosen to bound the *Family-Wise Error Rate (FWER)*, that is, the probability that any of the returned pairs be a spurious discovery, or to bound the *False Discovery Rate (FDR)*, that is, the fraction of returned pairs that can be spurious discoveries. The authors compare the effectiveness of the two approaches, but do not provide support thresholds in order to confine the analysis to smaller subsets of itemsets, thus making the mining task computationally feasible.

Beyond frequent itemset mining, the issue of significance has also been addressed in the realm of discovering association rules. In [12], the authors provide a variation of the well-known Apriori strategy for the efficient discovery of the set $\mathcal{A}$ of all association rules with p-value below a given cutoff value, while the results in [16] provide the means of evaluating the FDR in $\mathcal{A}$. The FDR metric is also employed in [24] in the context of discoverying significant quantitative rules, a variation of association rules. Again, these works do not address the crucial problem of determining an appropriate cutoff value that yields a reasonable number of discoveries while limiting the number of spurious discoveries.

1.2 Our Results

In this paper we address the classical problem of mining frequent itemsets with respect to a certain minimum support threshold, but provide a rigourous methodology to establish such a threshold so to guarantee, in a statistical sense, that the returned family of frequent itemsets contains significant ones with a limited false discovery rate. Our methodology crucially relies on the following Possion approximation result which is the main theoretical contribution of the paper.

Consider a dataset $\mathcal{D}$ of t transactions, on a set $\mathcal{I}$ of n items, where each transaction is a subset of $\mathcal{I}$. Let $\hat{\mathcal{D}}$ be a random dataset of t transactions on the same set of items $\mathcal{I}$, in which the occurrences of individual items in a transaction are independent, and the probability that a given item is included in a transaction is set to be its frequency in $\mathcal{D}$. (This is the same random model of [20].) Let $Q_{k,s}$ be the number of itemsets of size k that appear in *at least* s transactions of $\mathcal{D}$, and let $\hat{Q}_{k,s}$ be the corresponding random variable for $\hat{\mathcal{D}}$. We show that for certain meaningful ranges of parameters, the distribution of $\hat{Q}_{k,s}$ is well approximated by a Poisson distribution. The technique is based on a novel application of the Chen-Stein Poisson approximation method [3].

COMMENT: The following 2 paragraphs may need to be restructured based on which algorithms we decide to include.

Based on the above Poisson approximation and on techniques for limiting the FDR when testing multiple hypotheses, we set up a methodology for determining a minimum

support threshold for the mining task. Specifically, for a fixed k we test a small set of values of the support threshold s in the range where the Poisson approximation holds, measuring the p-value corresponding to the null hypothesis H_0 that the observed value $Q_{k,s}$ comes from a Poisson distribution of suitable expectation. We then use the technique by [4] to determine a subset of supports for which H_0 is rejected and such that the fraction of rejected true hypotheses (FDR) is small. Among the identified supports, we pick the smallest one as the threshold for the mining task.

It is important to remark that the Poisson approximation holds for a range of supports where the random variable $\hat{Q}_{k,s}$ has low expectation, that is, a range of supports where the occurence of itemsets of size k can be regarded as a relatively "rare" event. As a consequence, the support threshold returned by our strategy is likely to yield a reasonable number frequent itemsets of high significance, with few false discoveries, which is a very desirable feature for the computational feasibility of the mining task.

In order to complement our theoretical findings, in the last part of the paper we report on a number of experiments on several real datasets that are standard benchmarks for frequent itemsets mining.

COMMENT: Possible list of experiments to be reported:

- **For each dataset $\mathcal{D}$, we determine the minimum support value for which the Poisson approximation holds for the corresponding random dataset $\hat{\mathcal{D}}$**
- **We apply the algorithm(s) to the datasets and comment on the output possibly comparing the various algorithms. Remark here that for most (but not all) pairs (dataset, k) we obtain "few" significant itemsets and low FDR.**
- **We do a sanity check on the corresponding random datasets to see that our methodology returns (almost) nothing**
- **Chernoff bound: we use it as a witness for any methodology concentrating on single itemsets to show that if we want to keep the FWER low, a large nummber of false negatives (hence many undiscovered signficant itemsets) have to be expected.**

1.3 Organization of the Paper

The rest of the paper is structured as follows. In Section 2, we illustrate some backgroung material on multiple hypotheses testing and introduce the benchmark datasets that will be used in the experiments. Section 3 describes a technique to obtain significant itemsets with small FWER. Section 4 contains the proof of the Poisson approximation for the random variable $\hat{Q}_{k,s}$. The ensuing methodology for determining suitable support thresholds for mining significant itemsets is illustrated in Section 5, while Section 6 reports the experimental results. Section 7 offers some closing remarks.

Dataset	n	$[f_{\min}; f_{\max}]$	m	t
Retail	16470	[1.13e-05 ; 0.57]	10.3	88162
Kosarac	41270	[1.01e-06 ; 0.61]	8.1	990002
Bms1	497	[1.68e-05 ; 0.06]	2.5	59602
Bms2	3340	[1.29e-05 ; 0.05]	5.6	77512
Bmspos	1657	[1.94e-06 ; 0.60]	7.5	515597
Pumsb*	2088	[2.04e-05 ; 0.79]	50.5	49046

Table 1: Parameters of the benchmark datasets: n is the number of items; $[f_{\min}, f_{\max}]$ is the range of frequencies of the individual items; m is the average transaction length; and t is the number of transactions.

2. Preliminaries

2.1 Multiple hypothesis testing

COMMENT: the following points need to be expanded

- **Define multiple hypothesis testing**
- **Define FWER and FDR and argue that the former suffers from too many false negatives**
- **Present Benjamini-Hochberg technique to limit FDR**

2.2 Benchmark datasets

In order to validate the methodology, a number of experiments, whose results are reported in Section 6, have been performed on datasets which are standard benchmarks in the context of frequent itemsets mining. The main characteristics of the datasets we use are summarized in Table 1. A description of the datasets can be found in the FIMI Repository (`http://fimi.cs.helsinki.fi/data/`), where they can also be downloaded.

3. Obtaining significant itemsets with small FWER

In this section, we introduce an itemset-wise strategy for the discovery of significant itemsets with a limited FWER, which however, may result in a lot of false negatives and cannot be easily employed to establish a suitable minimum support threshold. As before, let $\mathcal{D}$ be a dataset of t transactions over a set $\mathcal{I}$ of n items, and $\hat{\mathcal{D}}$ be the corresponding random dataset with the same number of transactions and the same frequencies of individual items. We wish to flag as significant any itemset $X \subseteq \mathcal{I}$ for which the dependencies among its constituent items are such that the support of X in $\mathcal{D}$ is substantially larger than its expected support in $\hat{\mathcal{D}}$. The following theorem provides a means of identifying such a large deviation.

Theorem 1. *Let k be an integer, with $1 < k \leq n$, and α a user-defined parameter, with $0 < \alpha < 1$. For each itemset $X \subseteq \mathcal{I}$ of size k let q_X be the product of the frequencies of*

the items in X, and $\mu_X = q_X \cdot t$ the expected support of X in $\hat{\mathcal{D}}$. Define the support threshold $\bar{s}(X)$ as[2]*:*

$$\bar{s}(X) = \begin{cases} (1+\epsilon)\mu_X & \exists \epsilon < 1 : \epsilon^2 \mu_X \geq (3/\log e)\log(n^k/\alpha) \\ \max[6\mu_X, \log(n^k/\alpha)] & \textit{otherwise} \end{cases}$$

The probability that $\hat{\mathcal{D}}$ contains any itemset X of size k with support greater than or equal to $\bar{s}(X)$ is less than α.

PROOF. Let $s(X)$ denote the support of set X in the random dataset $\hat{\mathcal{D}}$. It is immediate to see that $s(X)$ is a binomial random variable with parameters t and q_X and expectation μ_X. The support threshold $\bar{s}(X)$ is chosen so that by applying the Chernoff bound [17, Theorem 4.4] we have that

$$\mathbf{Pr}(s(X) \geq \bar{s}(X)) \leq \frac{\alpha}{\binom{n}{k}}.$$

Thus,

$$\sum_{X:|X|=k} \mathbf{Pr}(s(X) \geq \bar{s}(X)) \leq \alpha$$

□

Suppose that we apply the above theorem to the test all itemsets of size k, flagging as significant those whose observed supports in $\mathcal{D}$ are higher than the respective specified thresholds. It is easy to argue that α provides an upper bound to the FWER of the family of significant itemsets obtained in this fashion. Unfortunately, as mentioned above this approach suffers from two main drawbacks. First, testing *all itemsets* of a given size k quickly becomes unfeasible as n and k grow. Second, the FWER metric is too stringent, hence many dependencies among items may not be discovered (false negatives) [5]. (A concrete evidence of the latter phenomenon will be presented in Section 6.)

Next, we describe a methodology that identifies, for each k, a support threshold s so that all itemsets of size k and support greater than or equal to s can be flagged as significant with a small FDR. The theoretical foundation of our approach rests on a new Poisson approximation result, discussed in the following section.

4. Possion Approximation for $\hat{Q}_{k,s}$

The Chen-Stein method [3] is a powerful tool for bounding the error in approximating probabilities associated with a sequence of dependent events by a Poisson distribution. To apply the method to our case, we fix parameters k and s, and define a collection of Bernoulli random variables $\{Z_X \mid X \subset \mathcal{I},\ |X| = k\}$, such that $Z_X = 1$ if the itemset X appears in at least s transactions in the random dataset $\hat{\mathcal{D}}$, and $Z_X = 0$ otherwise. Also, let $p_X = \mathbf{Pr}(Z_X = 1)$. We are interested in the distribution of $\hat{Q}_{k,s} = \sum_{X:|X|=k} Z_X$.

[2]All logarithms in the paper are taken to the base two.

For each set X we define the *neighborhood set* of X,

$$I(X) = \{X' \mid X \cap X' \neq \emptyset, |X'| = |X|\}.$$

If $Y \notin I(X)$ then Z_Y and Z_X are independent. Adapting [3, Theorem 1] to our case we have:

Theorem 2. *Let U be a Poisson random variable such that $\mathbf{E}[U] = \mathbf{E}[\hat{Q}_{k,s}] = \lambda < \infty$. The variation distance between the distributions $\mathcal{L}(\hat{Q}_{k,s})$ of $\hat{Q}_{k,s}$ and $\mathcal{L}(U)$ of U is such that*

$$\left\| \mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U) \right\| = \sup_A |\mathbf{Pr}(\hat{Q}_{k,s} \in A) - \mathbf{Pr}(U \in A)| \leq b_1 + b_2,$$

where

$$b_1 = \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y$$

and

$$b_2 = \sum_{X:|X|=k} \sum_{X \neq Y \in I(X)} \mathbf{E}[Z_X Z_Y].$$

COMMENT: don't we need a factor 2 multiplying $\sup_A \cdots$ and $b_1 + b_2$? See [3, Sect.3]. This factor is also missing in the proof of Theorem 3

It is easy to see that the quantities b_1 and b_2 in the above theorem are both decreasing in s. Therefore, if $b_1 + b_2 < \epsilon$ for a given s, then the same upper bound will hold for every $s' > s$. Consequently, a given Poisson approximation for $\hat{Q}_{k,s}$, established thorough the above theorem, extends to $\hat{Q}_{k,s'}$ with $s' > s$.

We can compute explicit bounds for b_1 and b_2 in many situations. Specifically, suppose that we generate t transactions in the following way. For each item x, we sample a random variable $R_x \in [0,1]$ independently from some distribution R. Conditioned on the R_x's, each item x occurs independently in each transaction with probability R_x. In what follows, we provide specific bounds for this situation that depend on the moment $\mathbf{E}[R^{2s}]$ of the random variable R.

COMMENT 1: We think we should rephrase Theorem 3 to make explicit the fact that the Poisson approximation holds for a range of supports $s \geq s_{\min}$ (see the new paragraph added after Theorem 2).

COMMENT 2: the theorem below aims at $b_1 + b_2 = O(1/n)$ but our experiments aim at $b_1 + b_2 < 0.01$. Shall we make these aims consistent?

Theorem 3. *Consider an asymptotic regime where as $n \to \infty$, we have $k, s = O(1)$ with $s \geq 2$, $\mathbf{E}[R^{2s}] = O(n^{-a})$ for some constant $2 \leq a \leq 2s$, and $t = O(n^c)$ for some constant $0 < c < (a-2)(k-1)/2s$. If*

$$c \leq \frac{k(a-2) + \max(a-4, 0)}{2s},$$

then the variation distance between the distributions

$\mathcal{L}(\hat{Q}_{k,s})$ and $\mathcal{L}(U)$ of $\hat{Q}_{k,s}$ and U satisfies

$$\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| = \sup_A |\mathbf{Pr}(\hat{Q}_{k,s} \in A) - \mathbf{Pr}(U \in A)| = O(1/n).$$

PROOF. Applying Theorem 2 gives

$$\left\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\right\| \leq b_1 + b_2$$

where

$$b_1 = \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y$$

and

$$b_2 = \sum_{X:|X|=k} \sum_{Y \neq X \in I(X)} \mathbf{E}[Z_X Z_Y].$$

We now evaluate b_1 and b_2. Letting $\vec{R}$ denote the vector of the R_x's, we have that for any set X of k items

$$\mathbf{Pr}(Z_X = 1 \mid \vec{R}) \leq \binom{t}{s} \prod_{x \in X} R_x^s.$$

Since the R_x's are independent with common distribution R,

$$p_X = \mathbf{E}[\mathbf{Pr}(Z_X = 1 \mid \vec{R})] \leq \binom{t}{s} \mathbf{E}[R^s]^k.$$

Using Jensen's inequality, we now have

$$\begin{aligned}
b_1 &= \sum_{X:|X|=k} \sum_{Y \in I(X)} p_X p_Y \\
&\leq \left(\binom{n}{k}^2 - \binom{n}{k}\binom{n-k}{k}\right)\binom{t}{s}^2 \mathbf{E}[R^s]^{2k} \\
&\leq \left(\binom{n}{k}^2 - \binom{n}{k}\binom{n-k}{k}\right)\binom{t}{s}^2 \mathbf{E}[R^{2s}]^{k} \\
&= \binom{n}{k}^2 \left(1 - \frac{\binom{n-k}{k}}{\binom{n}{k}}\right)\binom{t}{s}^2 \mathbf{E}[R^{2s}]^{k} \\
&= \binom{n}{k}^2 \left(1 - \prod_{i=0}^{k-1} \frac{n-k-i}{n-i}\right)\binom{t}{s}^2 \mathbf{E}[R^{2s}]^{k} \\
&= \Theta(n^k)^2 \cdot \Theta(1/n) \cdot O(n^{2cs}) \cdot O(n^{-ka}) \\
&= O(n^{k(2-a)+2cs-1})
\end{aligned}$$

We now turn our attention to b_2. Consider sets $X \neq Y$ of k items, let $g = |X \cap Y|$, and suppose that $g > 0$. Then if $Z_X Z_Y = 1$, there exist disjoint subsets $A, B, C \in \{1, \ldots, t\}$ such that $0 \leq |A| \leq s$, $|B| = |C| = s - |A|$, all of the transactions in A contain both X and Y, all of the transactions in B contain X, and all of the transactions in C contain Y. Therefore,

$$\begin{aligned}
\mathbf{E}[Z_X Z_Y \mid \vec{R}] &\leq \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\prod_{x \in X \cup Y} R_x^i\right) \times \\
&\quad \times \left(\prod_{x \in X} R_x^{s-i}\right)\left(\prod_{y \in Y} R_y^{s-i}\right) \\
&= \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \left(\prod_{x \in X \cap Y} R_x^{2s-i}\right) \times \\
&\quad \times \left(\prod_{x \in X - Y} R_x^{s}\right)\left(\prod_{y \in Y - X} R_y^{s}\right).
\end{aligned}$$

Applying independence of the R_x's and Jensen's inequality gives

$$\begin{aligned}
\mathbf{E}[Z_X Z_Y] &= \mathbf{E}[\mathbf{E}[Z_X Z_Y \mid \vec{R}]] \\
&\leq \sum_{i=0}^{s} \binom{t}{i; s-i; s-i} \mathbf{E}[R^{2s-i}]^g \mathbf{E}[R^s]^{2(k-g)} \\
&\leq \sum_{i=0}^{s} t^{2s-i} \mathbf{E}[R^{2s}]^{\frac{g(2s-i)}{2s}} \mathbf{E}[R^{2s}]^{k-g} \\
&= \sum_{i=0}^{s} t^{2s-i} \mathbf{E}[R^{2s}]^{k-ig/2s} \\
&\leq O(1) \sum_{i=0}^{s} n^{(2s-i)c - a(k-ig/2s)} \\
&= O(1) \sum_{i=0}^{s} n^{2sc-ak+i\left(\frac{ag}{2s}-c\right)} \\
&= O(n^{2sc-ak}) \sum_{i=0}^{s} n^{i\left(\frac{ag}{2s}-c\right)} \\
&= O(n^{2sc-ak}) \begin{cases} \Theta(1) & g \leq 2sc/a \\ \Theta(n^{s\left(\frac{ag}{2s}-c\right)}) & \text{otherwise} \end{cases}
\end{aligned}$$

It follows that

$$\begin{aligned}
b_2 &\leq \sum_{g=1}^{k-1} \binom{n}{g; k-g; k-g} O(n^{2sc-ak}) \begin{cases} \Theta(1) & g \leq 2sc/a \\ \Theta(n^{s\left(\frac{ag}{2s}-c\right)}) & \text{otherwise} \end{cases} \\
&\leq O(n^{2k-g+2sc-ak}) \begin{cases} \Theta(1) & g \leq 2sc/a \\ \Theta(n^{s\left(\frac{ag}{2s}-c\right)}) & \text{otherwise} \end{cases} \\
&= O(n^{2k+2sc-ak}) \sum_{g=1}^{k-1} n^{-g} \begin{cases} \Theta(1) & g \leq 2sc/a \\ \Theta(n^{s\left(\frac{ag}{2s}-c\right)}) & \text{otherwise} \end{cases}.
\end{aligned}$$

Now, for $2sc/a < g < k$, we have (using the fact that $a \geq 2$)

$$n^{-g} n^{s\left(\frac{ag}{2s}-c\right)} = n^{g\left(\frac{a}{2}-1\right)-cs} \leq n^{(k-1)\left(\frac{a}{2}-1\right)-cs}.$$

Thus, $b_2 = O(n^{2k+sc-ak+(k-1)\left(\frac{a}{2}-1\right)})$. (Here we are using the assumption that $c \leq (k-1)(a-2)/2s$ to ensure that $n^{(k-1)\left(\frac{a}{2}-1\right)-cs} = \Omega(1)$.)

Dataset	$s_{\min}$		
	$k=2$	$k=3$	$k=4$
RandRetail	9237	4366	784
RandKosarak	273266	100543	20120
RandBms1	268	23	5
RandBms2	168	13	4
RandBmspos	76672	15714	2717
RandPumsb*	29303	21893	16265

Table 2: Minimum support $s_{\min}$ for which the Poisson approximation for $\hat{Q}_{k,s}$ holds (i.e., $b_1 + b_2 < 0.01$) for $k = 2, 3, 4$ in random datasets with the same values of n, t, m and with the same frequencies of the items as the corresponding benchmark datasets.

Now, we have $b_1 = O(1/n)$ if $c \leq k(a-2)/2s$ and $b_2 = O(1/n)$ if $c \leq [k(a-2)+(a-4)]/2s$. Thus, $b_1+b_2 = O(1/n)$ if

$$c \leq \frac{k(a-2) + \max(a-4, 0)}{2s}.$$

□

The above theorem provides rigorous analytical evidence that there exists a meaningful range for the support threshold s such that the number of itemsets of size k with that support or larger can be approximated by a Poisson variable. In practice, in order to avoid the inevitable slack due to the use of asymptotics in the theorem, the range of validity of the Poisson approximation can be conveniently established via simple Montecarlo simulations to estimate the values b_1 and b_2 of Theorem 2. As an application, for each dataset $\mathcal{D}$ of Table 1 and for itemset sizes $k = 2, 3, 4$, we determined the minimum value $s_{\min}$ so that the sum $b_1 + b_2$ is below 0.01 for a corresponding random dataset $\hat{\mathcal{D}}$. The values of $s_{\min}$ we obtained are reported in Table 2 (we added the prefix "Rand" to each dataset name, to denote the fact that the dataset is random and features the same parameters as the corresponding real one).

COMMENT: We should probably give more details about the monte carlo simulations

5. Obtaining a Support Threshold for Mining Significant Itemsets

COMMENT: This section should contain a 'complete' recipe for mining significant itemsets. The recipe could be structured in 2 phases: (1) find the minimum support $s_{\min}$ such that the Poisson approximation holds (either applying the theorem or running the simulations as explained in the previous secton); (2) find the threshold $s*$ and use it for mining. Here we have to decide how many algorithms to present for Phase 2. We think that if we present more than one algorithm we should have some clear evidence (possibly experimental) of the advantages of each algorithm.

6. Experimental Results

Possible list of experiments to report:

- **Results of the application of the "recipe" to the benchmark datasets. Features to stress: low FDR; $s*$ identifies a region of rare events hence likely to contain few very significant itemsets; for some pairs (dataset, k) the null hypothesis is not rejected, which is evidence that for that particular k the dataset is likely to behave like a random one.**
- **Results of a sanity check on random datasets showing that our methodology returns (almost) nothing**
- **Results of the application of the technique of Section 3 on the itemsets mined with threshold $s*$ given by the Poisson method, to show that that extreme techniques suffers from a non-negligible number of false negatives.**

7. Conclusions

8. References

[1] C.C. Aggarwal and P.S. Yu. A new framework for itemset generation. In *Proc. of the 17th ACM Symp. on Principles of Database Systems*, pages 18–24, 1998.

[2] R. Agrawal, T. Imielinski, and A.N. Swami. Mining association rules between sets of items in large databases. In *Proc. of the ACM SIGMOD Intl. Conference on Management of Data*, pages 207–216, 1993.

[3] R. Arratia, L. Goldstein, and L. Gordon. Poisson approximation and the Chen-Stein method. *Statistical Science*, 5(4):403–434, 1990.

[4] Y. Benjamini, and Y. Hochberg. Controlling the false discovery rate. *J. Royal Statistical Society*, Series B, 57:289–300, 1995.

[5] R.J. Bolton, D.J. Hand, and N.M. Adams. Determining Hit Rate in Pattern Search In *Proc. of Pattern Detection and Discovery*, LNAI 2447, pages 36–48, 2002.

[6] W. J. Conover. *Practical Nonparametric Statistics*. Wiley Series in Probability, 3rd Ed., 1999.

[7] W. DuMouchel. Bayesian data mining in large frequency tables, with an application to the FDA spontaneous reporting system. *The American Statistician*, 53:177–202, 1999.

[8] W. DuMouchel and D. Pregibon. Empirical Bayes screening for multi-item associations. In *Proc. of the 7th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, pages 67–76, 2001.

[9] A. Gionis, H. Mannila, T. Mielikäinen, and P. Tsaparas. Assessing data mining results via swap randomization. In *Proc. of the 12th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, pages 167–176, 2006.

[10] B. Goethals, R. Bayardo, and M. J. Zaki, editors. *Proc. of the 2nd Workshop on Frequent Itemset Mining Implementations (FIMI04)*, volume 126. CEUR-WS Workshop On-line Proceedings, November 2004.

[11] B. Goethals and M. J. Zaki, editors. *Proc. of the 1st Workshop on Frequent Itemset Mining Implementations (FIMI03)*, volume 90. CEUR-WS Workshop On-line Proceedings, November 2003.

[12] W. Hämäläinen, and M. Nykänen Efficient discovery of statistically significant association rules In *Proc. of the 8th IEEE Intl. Conference on Data Mining*, 2008. To appear.

[13] J. Han, H. Cheng, D. Xin, and X. Yan. Frequent pattern mining: Current status and future directions. *Data Mining and Knowledge Discovery*, 14(1), 2007.

[14] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, San Mateo, CA, 2001.

[15] H.O. Lancaster. *The Chi-squared Distribution*. John Wiley & Sons, New York NY, 1969.

[16] N. Megiddo, and R. Srikant Discovering predictive association rules. In *Proc. of the 4th Intl. Conference on Knowledge Discovery and Data Mining*, pages 274–278, 1998.

[17] M. Mitzenmacher and E. Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis.* Cambridge University Press, 2005.

[18] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In *Proc. of the 7th Int. Conference on Database Theory*, pages 398–416, January 1999.

[19] A. Silberschatz and A. Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Trans. on Knowledge and Data Engineering*, 8(6):970–974, 1996.

[20] C. Silverstein, S. Brin, and R. Motwani. Beyond market baskets: Generalizing association rules to dependence rules. *Data Mining and Knowledge Discovery*, 2(1):39–68, 1998.

[21] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proc. of the ACM SIGMOD Intl. Conference on Management of Data*, pages 1–12, 1996.

[22] P. Tan, M. Steinbach, and V. Kumar. *Introduction to Data Mining*. Addison Wesley, 2006.

[23] D. Xin, J. Han, X. Yan, and H. Cheng. Mining compressed frequent-pattern sets. In *Proc. of the 31st Very Large Data Base Conference*, pages 709–720, 2005.

[24] H. Zhang, B. Padmanabhan, and A. Tuzhilin. On the discovery of significant statistical quantiative rules. In *Proc. of the 10th ACM SIGKDD Intl. Conference on Knowledge Discovery and Data Mining*, pages 374–383, 2004.