Probabilistic analysis of success and failure rates of candidates generation algorithms for the frequent itemsets mining problem

Minh Tang
mhtang@cs.indiana.edu
Department of Computer Science
Indiana University Bloomington

Abstract

This paper consider the success and failure probability of candidate generation algorithms for the frequent itemsets mining problem under several probability model. Results for one of the models had been obtained previously, but with a complex derivation. Our re-derivation of these results is simpler and employed a concentration inequality for the sum of independent Bernoulli random variables. Our results for the other models employed concentration inequalities for martingales and is applicable to models where there is dependence between the transactions. From the success and failure probability we can estimate the size of the maximum frequent itemset.

1 Introduction

The frequent itemsets mining problem is the problem of given a set of items $\mathcal{I}$, a set of transactions $\mathcal{D}$, and a threshold value σ, find all subsets of $\mathcal{I}$ whose occurrences in $\mathcal{D}$ is at least σ. The mining of frequent itemsets is an important problem in many data mining tasks, especially in the context of finding association rules [1].

The class of candidates generation algorithms includes the Apriori algorithm [1] and the Eclat algorithm [13], two of the most widely used algorithm in frequent itemset mining. Since the running time of any frequent itemsets mining algorithm is exponential in the worst case, the average time analysis is more important in the understanding of the efficiency of the algorithms. Previous work had been done in [6] regarding the average time analysis of the Apriori and Eclat algorithms under the assumption that for every subset $J \subseteq \mathcal{I}$ of items, the probability that a transaction $T \in \mathcal{D}$ contains J is $P(J)$, independently of all other transactions $T' \in \mathcal{D}$. Although it's not explicitly computed in [6], the analysis done in [6] also allowed us to bound probabilistically the maximum size of the frequent itemsets, i.e. there exist a constant μ such that the probability that the maximum size of the frequent itemsets exceeds $\mu + \delta$ decreases exponentially with increasing $\delta > 0$. Given that in general, we cannot approximate the maximum size of the frequent itemsets in a set of n items within a factor of n^{ϵ} for some $\epsilon > 0$ unless $P = NP$ [3], the probabilistic bound in [6] is therefore very important in understanding the efficiency of the above mentioned algorithms.

Our paper is then composed of two parts. The first part is a re-derivation of the important results for the success and failure probability of the Apriori and Eclat algorithm in the vein of [6], but with a much simpler argument. The second part is the use of martingales to derive additional results for the success probabilities for the mentioned algorithms when the probability model doesn't assume independence between the transactions. The organization of the paper is then as follow. We introduce the Apriori algorithm and the associated probability model in Section 2. We present in Section 3 and 4 our analysis of the success and failure probabilities of the Apriori and Eclat algorithm, respectively, along with comparisons between our results and those obtained in [6]. Section 5 apply the bounds for the success probabilities to determine the size of the largest frequent itemset for two simple probability assignments. The application of martingales' concentration inequalities to the analysis of the success probabilities is investigated in Section 6.

2 Background

Let $\mathcal{I} = \{i_1, i_2, \ldots, i_n\}$ be a set of items. A set of transactions $\mathcal{T}$ is a family of subsets of $\mathcal{I}$. The support of a subset $I \subseteq \mathcal{I}$ with respect to a set of transactions $\mathcal{T}$, $\mathrm{supp}(I, \mathcal{T})$, is the number of transactions $T \in \mathcal{T}$ such that $I \subseteq T$. The subset $I \subseteq \mathcal{I}$ is frequent in $\mathcal{T}$ if $\mathrm{supp}(I, \mathcal{T}) \geq \sigma$ for some threshold value $\sigma > 0$. The frequent set mining problem is then, given the set of items $\mathcal{I}$ and the set of transactions $\mathcal{T}$, find all $I \subseteq \mathcal{I}$ such that I is frequent with respect to $\mathcal{T}$.

The Apriori algorithm for mining frequent itemsets is a breadth-first, level-wise algorithm where for each level k, the Apriori algorithm first prunes all infrequent itemsets of size k and then generates the candidate itemsets of size $k + 1$. A set J of size $k + 1$ is a candidate if and only if all of its subset of size k is frequent, i.e. not pruned by the Apriori algorithm at level k. The algorithm stops when no other frequent itemsets can be found.

We say that a candidate set J is a success if J is also frequent, otherwise, J is a failure. Since the support of J in $\mathcal{T}$ is computed for any candidate set J, we would prefer that J is also a success. The success and failure probabilities of the candidate sets are therefore important in the analysis of the average running time of the Apriori algorithm. If the failure probabilities of the candidate sets are small, then almost all of the computational work done by the Apriori algorithm are necessary. Otherwise, if the success probabilities of the candidate sets are small, then a lot of the computational work done are in some sense useless.

The analysis that we will be doing in this paper is then the analysis of the success and failure probabilities of the candidate sets. Here we discussed briefly the model and approach used in [6]. Given a set $I \subseteq \mathcal{I}$, the probability that a transaction $T \in \mathcal{T}$ satisfy $T \supseteq I$ is denoted by $P(I)$. We also assume that the transactions in $\mathcal{T}$ are independent. Then for a given I, we can view each transactions $T \in \mathcal{T}$ as a Bernoulli trial with $p = P(I)$ and $q = 1 - P(I)$. The success probability $S(I)$ can then be written as ([6], Eq. 3.2)

$$S(I) = \sum_{j \geq \sigma} \binom{b}{j} [P(I)]^j [1 - P(I)]^{b-j}$$

Let's assume that $I = \{i_1, i_2, \ldots, i_m\}$ is of size m and denote by $I_1, I_2, \ldots, I_m$ all the m subsets of I of size $m - 1$. I is a candidate frequent set during the execution of the Apriori algorithm if the subsets $I_1, I_2, \ldots I_m$ are all frequent. If we let j_0 be the number of transactions containing all items in I and j_k, $1 \leq k \leq m$, be the number of transactions containing all items of I_k without containing all items of I, then I is a candidate if

$$j_0 + j_1 \geq \sigma \wedge j_0 + j_2 \geq \sigma \wedge \cdots \wedge j_0 + j_m \geq \sigma \tag{2.1}$$

For ease of notation, let $I_0 = I$, and define $Q(I_k)$ for $1 \leq k \leq m + 1$ as follows

$$Q(I_0) = P(I), \quad Q(I_k) = P(I_k) - P(I) \;\; 1 \leq k \leq m, \quad Q(I_{m+1}) = 1 - \sum_{k=0}^{m} Q(I_k)$$

So $Q(I_k)$ for $1 \leq k \leq m$ is the probability that a transaction contains I_k but not I while $Q(I_{m+1})$ is the probability that a transaction doesn't contain any of the I_k for $1 \leq k \leq m$. Finally, let $j_{m+1} = b - j_0 - j_1 - \cdots - j_m$. The candidate probability $C(I)$

and failure probability $F(I)$ can then be written as ([6], Eq. 3.35)

$$C(I) = \sum_{\substack{j_0 \\ j_1 \geq \sigma - j_0 \\ j_2 \geq \sigma - j_0 \\ \vdots \\ j_m \geq \sigma - j_0}} \binom{b}{j_0, j_1, \ldots, j_m, j_{m+1}} \prod_{k=0}^{m+1} Q(I_k)^{j_k} \tag{2.2}$$

$$F(I) = \sum_{\substack{j_0 < \sigma \\ j_1 \geq \sigma - j_0 \\ j_2 \geq \sigma - j_0 \\ \vdots \\ j_m \geq \sigma - j_0}} \binom{b}{j_0, j_1, \ldots, j_m, j_{m+1}} \prod_{k=0}^{m+1} Q(I_k)^{j_k} \tag{2.3}$$

The results in [6] for the Apriori algorithm were derived by applying the Chernoff bounds [4] to Eq. (2.1) through Eq. (2.3). We will explicitly state those results in the following section.

3 Success & failure probabilities: Apriori algorithm

Our derivation of the success and failure probability of the Apriori algorithm is based on a different approach than that of [6] which was described in the previous section. Let X_I be a random variable with $X_I = |\{T \in \mathcal{T} : I \subseteq T\}|$. We see that X_I had a binomial distribution with probability $p = P(I)$ and $n = |\mathcal{T}| = b$.Given a set I, the characteristic function $\xi_I : \mathcal{T} \mapsto \{0, 1\}$ is defined as

$$\xi_I(T) = \begin{cases} 1 & \text{if } T \supseteq I \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

The set I is therefore a success if $\sum_{T \in \mathcal{T}} \xi_I(T) = X_I \geq \sigma$. If $S(I)$ is the success probability of the set I, then from the above reasoning,

$$S(I) = \Pr\left(\sum_{T \in \mathcal{T}} \xi_I(T) \geq \sigma\right) \tag{3.2}$$

The following inequality on the concentration of the sum of independent random variables is essential to our analysis [9].

Theorem 3.1: *Let $\xi_1, \xi_2, \ldots, \xi_n$ be independent random variables, $X_k - \mathbb{E}[X_k] \leq \eta$ for all k. We consider the sum $X = \xi_i + \xi_2 + \cdots + \xi_n$, with mean μ and variance V. Then for any $t \geq 0$*

$$\Pr(X - \mu \geq \lambda) \leq e^{-\frac{\lambda^2}{2V(1+\eta t/3V)}} \tag{3.3}$$

Since the set of transactions are assumed to be independent, given a subset $I \subseteq \mathcal{I}$, the expected number of transactions containing I is $\mathbb{E}(X_I) = bP(I)$ where $b = |\mathcal{T}|$ is the number of transactions. Suppose that $bP(I) \le \sigma$, then the success probability $S(I)$ can be rewritten as

$$S(I) = \Pr(X_I \ge \sigma) = \Pr(X_I \ge bP(I) + \lambda)$$

where $\lambda = \sigma - bP(I)$. Each of the ξ_k is a Bernoulli random variable with mean $P(I)$ and variance $P(I)(1 - P(I))$ so $\mu = bP(I)$ and $V = bP(I)(1 - P(I))$. In addition, $\eta = \max(P(I), 1 - P(I))$. If $\eta = o(b)$, then by Eq. (3.3), we have

$$S(I) = \Pr(X_I \ge bP(I) + \lambda) \le e^{-\frac{\lambda^2}{2V(1+\eta\lambda/3V)}} \le e^{-\frac{\lambda^2}{2V}(1+o(1))} \tag{3.4}$$

We will now state the corresponding result, Theorem 3.4.1 from [6].

(Theorem A): *When $P(I) \le \sigma/b$, the following upper bound for $S(I)$ can be found*

$$S(I) \le e^{-b\alpha_1^2/\{2P(I)(1-P(I))\}+O(b\alpha_1^3(1-P(I))^{-2})} \tag{3.5}$$

where $\alpha_1 = \sigma/b - P(I)$. In this case $S(I)$ goes to 0 rapidly with increasing α_1.

If we ignore the less significant terms, the $o(1)$ term in (3.4) and the $O(b\alpha_1^3)$ term in Eq. (3.5), then we see that their bounds are in fact equivalent. In both cases, the success probability $S(I)$ decreases rapidly towards 0 as λ increases.

On the other hand, if $bP(I) \ge \sigma$, then the success probability can be written as

$$S(I) = \Pr(X_I \ge \sigma) = 1 - \Pr(X_I \le \sigma - 1) = 1 - \Pr(X_I \le bP(I) - \lambda)$$

where $\lambda = bP(I) - (\sigma - 1)$. We cannot use Eq. (3.3) directly since the inequality in Eq. (3.3) is only one-sided. However, the random variables ξ'_k with $\xi'_k = 1 - \xi_k$ has mean $1 - P(I)$ and variance $P(I)(1 - P(I))$. Now,

$$\Pr(X \le \mu - \lambda) = \Pr(b - X \ge b - (\mu - \lambda)) = \Pr(Y \ge \mu' + \lambda)$$

where $Y = \sum \xi'_k = \sum (1 - \xi_k)$ and $\mu' = \mathbb{E}[Y]$. Therefore, since the variance of Y is identical to the variance of X, we have

$$\Pr(X \le \mu - \lambda) \le e^{\frac{-\lambda^2}{2V}(1+o(1))} \tag{3.6}$$

The success probability $S(I)$ is then

$$S(I) \ge 1 - \Pr(X \le \mu - \lambda) \ge 1 - e^{\frac{-\lambda^2}{2V}(1+o(1))} \tag{3.7}$$

The corresponding result from [6] is Theorem 3.4.2

(Theorem B): *When $P(I) \geq (\sigma - 1)/b$, we have the following lower bound*

$$S(I) \geq 1 - e^{-b\alpha_2^2/\{2P(I)[1-P(I)]\}+O(b\alpha_2^3 P(I)^{-2})} \tag{3.8}$$

with $\alpha_2 = P(I) - (\sigma - 1)/b$. In this case, $S(I)$ goes to 1 rapidly with increasing α_2.

Once again, we see that the bounds given by Eq. (3.7) and Eq. (3.8) are equivalent.

To analyze the bounds for the failure probability $F(I)$ of a candidate set I, we first analyze the bounds for the candidacy probability $C(I)$ of I. Since $C(I) = S(I) + F(I)$, the bounds for the failure probability will follows from the bounds for $C(I)$ and $S(I)$. Let I be the set of items under considerations where $|I| = m$. Let $I_1, I_2, \ldots, I_m$ be the subsets of I of size $m - 1$. The set I is a candidate if all the I_j, $1 \leq j \leq m$ are frequent. We therefore have the following simple bounds on $C(I)$.

$$S(I_{\min})^m \leq C(I) \leq S(I_{\min}) \leq S(I_{\max}) \tag{3.9}$$

where $I_{\max}$ and $I_{\min}$ are subsets of I, $S(I_{\max}) \geq S(I_j) \geq S(I_{\min})$ for all $1 \leq j \leq m$.

Now, if $bP(I_j) < \sigma$ for any I_J, then by Eq. (3.4), $S(I_J)$ approaches 0 rapidly as $\lambda_j = \sigma - bP(I_j)$ increases. Therefore, Eq. (3.9) indicates that $C(I)$ approaches 0 rapidly as λ_j increases. Since $C(I)$ approaches 0, $F(I)$ also approaches 0 rapidly as λ_j increases. In short, we have

$$F(I) \leq C(I) \leq S(I_j) = \Pr(X_{I_j} \geq bP(I_j) + \lambda_j) \leq e^{-\lambda_j^2(1+o(1))/2V_j} \tag{3.10}$$

where $V_j = bP(I_j)(1 - P(I_j))$ is the variance of X_{I_j}. Of course, we want to chose the set I_j whose $\lambda_j = \sigma - bP(I_j)$ is maximum. We state the corresponding result, Theorem 3.5.1 from [6]

(Theorem C): *Let $Q(I_k)$ be $P(I_k) - P(I)$ for $1 \leq k \leq m$ and $Q(I_{\min})$ be the minimum values of the $Q(I_k)$. Also let l be the number of indices k such that $Q(I_k) = Q(I_{\min})$. In the region $Q(I) + P(I) \leq \sigma/b$, $F(I)$ goes rapidly to 0. In particular, when $\alpha_3 = \sigma/b - [Q(I) + P(I)]$*

$$F(I) \leq e^{-bl\theta\alpha_3^2/(2\{Q(I)+P(I)+(l-1)P(I)-l[Q(I)+P(I)]^2\})} \tag{3.11}$$

where θ is used to represent a function that approaches 1 in the limit.

The bound in Eq. (3.10) is in reality a bound for $C(I)$ rather then a bound on $F(I)$ while Eq. (3.11) is a bound for $F(I)$ directly. However, if we let v_1 and v_2 be the exponents in Eq. (3.10) and Eq. (3.11), respectively, then $v_1 = \vartheta v_2$ where $|\vartheta| \approx 1$.

Determining whether Eq. (3.10) or Eq. (3.11) gives a better bound seems to be tedious and unexciting, in our opinion, since in both cases, the failure probability and candidate probability both decrease to 0 rapidly as λ_j, or equivalently, α_3 increases.

If $bP(I) \geq \sigma$, then since $S(I)$ approaches 1, $C(I)$ also approaches 1 and $F(I)$ then approaches 0. In fact, with $C(I) \leq 1$, Eq. (3.7) gives

$$F(I) = C(I) - S(I) \leq 1 - (1 - e^{-\frac{\lambda^2}{2V}(1+o(1))}) \leq e^{-\frac{\lambda^2}{2V}(1+o(1))} \tag{3.12}$$

where $\lambda = bP(I) - (\sigma - 1)$. Once again, the bound in Eq. (3.12) coincides with the corresponding result, Theorem 3.5.2, from [6].

The only other case which we need to consider is if $bP(I) \leq \sigma \leq bP(I_j)$ for all I_j. From the definition of $S(I_j)$, we have

$$\begin{aligned} S(I_j) &= \Pr(X_{I_j} \geq \sigma) = 1 - \Pr(X_{I_j} \leq \sigma - 1) \\ &= 1 - \Pr(X_{I_j} - bP(I_j) \leq \lambda_j) \\ &\geq 1 - e^{-\lambda_j^2(1+o(1))/2V_j} \end{aligned} \tag{3.13}$$

where $\lambda_j = bP(I_j) - (\sigma - 1)$ and $V_j = bP(I_j)(1 - P(I_j))$ is the variance of X_{I_j}. Therefore, from the simple bound in Eq. (3.9), we have

$$C(I) \geq \left(1 - e^{-\lambda_{\min}^2(1+o(1))/2V_{\min}}\right)^m \tag{3.14}$$

where $\lambda_{\min}$ and $V_{\min}$ corresponds to the $S(I_{\min})$. We now use the following simple asymptotic estimate for the value of $(1-x)^k$

Proposition 3.2: *Let $0 < x < 1$. Then, for any positive integer k*

$$(1-x)^k \geq (1 - kx) \tag{3.15}$$

From Eq. (3.14) and Proposition 3.2, we have

$$C(I) \geq 1 - me^{-\lambda_{\min}^2(1+o(1))/2V_{\min}} \tag{3.16}$$

Therefore, for fixed m, $C(I) \to 1$ rapidly as $\lambda_{\min}$ increases. Since $bP(I) < \sigma$, by the Eq. (3.4), $S(I) \to 0$ as $\lambda = \sigma - bP(I)$ increases. Therefore, $F(I) \to 1$ as λ increases. In short,

$$F(I) = C(I) - S(I) \geq 1 - e^{-\frac{\lambda^2}{2V}(1+o(1))} - me^{-\lambda_{\min}^2(1+o(1))/2V_{\min}} \tag{3.17}$$

The corresponding result, Theorem 3.5.3 in [6] is

(Theorem D): *Let $\alpha_1 = \sigma/b - P(I)$ and $\beta_k = P(I_k) - (\sigma - 1)/b$ for $1 \le k \le m$ where α_1 and the β_k are all non-negative. Then we have the following lower bound*

$$\begin{aligned} F(I) \ge 1 &- e^{-b\alpha_1^2/\{2P(I)[1-P(I)]\}+O(b\alpha_1^3[1-P(I)]^{-2})} \\ &- \sum_{k=1}^{m} e^{-b\beta_k^2/\{2P(I_k)[1-P(I_k)]\}+O(b\beta_k^3[P(I_k)]^{-2})} \end{aligned} \tag{3.18}$$

Again, we see that the bound in Eq. (3.17) and Eq. (3.18) are very similar. In fact, the difference between the bound given by Eq. (3.17) and that given by Eq. (3.18) is the last term in the right hand side. The term in Eq. (3.18) is a sum over the β_k while Eq. (3.17) is m times the minimum term λ_{min}. We believe that the bound given by Eq. (3.18) might be slightly better, but in any case both bounds show that $F(I)$ increases to 1 rapidly as the λ_{min} and the α_1, β_k increases.

We now summarize the results of the preceding analysis. $I \subseteq \mathcal{I}$ is a set of m items.

1. If $bP(I) > \sigma$, then $S(I) \to 1$ rapidly as $\lambda = bP(I) - (\sigma - 1)$ increases. Since we have $C(I) = S(I) + F(I)$, $F(I) \to 0$ rapidly with increasing λ.
2. If $bP(I_j) < \sigma$ for any $I_j \subset I$ with $|I_j| = m - 1$, then $C(I) \to 0$ rapidly with increasing $\lambda_j = \sigma - bP(I_j)$. In this scenario, both $S(I)$ and $F(I)$ also approaches 0 rapidly with increasing λ_j.
3. If $bP(I_j) > \sigma > bP(I)$ for all $I_j \subset I$ with $|I_j| = m - 1$, then $C(I) \to 1$ rapidly as the λ_{min} of all the $\lambda_j = bP(I_j) - (\sigma - 1)$ increases. However since $\sigma > bP(I)$, $S(I) \to 0$ rapidly with increasing $\lambda = \sigma - bP(I)$. In this scenario, $F(I) \to 1$ rapidly as both λ_{min} and λ increases.

The above summary indicates that the number of transactions that contains the set I is concentrated around the mean $\mu = bP(I)$. But we can say a little more. From Eq. (3.4) and Eq. (3.7) we have

$$\Pr(|X_I - bP(I)| \ge t) \le 2e^{-\frac{t^2}{2V}(1+o(1))}$$

Now, $e^{-\frac{t^2}{2V}(1+o(1))}$ is small when $t^2 > 2V$, i.e. when $t = \Omega(\sqrt{2V}) = \Omega(\sqrt{bP(I)(1 - P(I))})$. The width of the concentration interval of X_I around $bP(I)$ is then $O(\sqrt{bP(I)(1 - P(I))})$. If $P(I)$ is not small enough, then the width of the concentration interval will be large and the resulting bounds for $S(I)$ and $F(I)$ will be weaker. Another equivalent statement to Eq. (3.4) and Eq. (3.7) is that

$$\Pr(|X_I - bP(I)| \ge \epsilon bP(I)) \le 2e^{-\epsilon^2 bP(I)/2(1-P(I))}$$

which indicates clearly that X_I is tightly concentrated around $\mu = bP(I)$.

4 Success and failure probabilities: Eclat algorithm

The Eclat algorithm [13] is another candidate generation algorithm for frequent itemset mining. The Eclat algorithm, contrary to the Apriori algorithm, is a depth-first search algorithm. In the Eclat algorithm, a set I is a candidate if two special subsets I_1 and I_2 of I is frequent. The set I_1 is the parent of I in the depth-first search tree, i.e. I_1 is the set I with the last augmented item removed. The set I_2 is the set I with the second-to-last augmented item removed. We will now see that the analysis of $S(I)$, $F(I)$, and $C(I)$ of the Apriori algorithm can be easily adapted to that of the Eclat algorithm.

The success probability $S(I)$ is straightforward. The analysis of $S(I)$ in Section 3 depends only the probability $P(I)$ and the number of transactions b in the database. Therefore, the success probability for the Eclat algorithm has identical bounds to that for the Apriori algorithm. The candidacy probability $C(I)$ is also straightforward. For the Eclat algorithm, Eq. (3.9) can be replaced by

$$\min(S(I_1), S(I_2))^2 \le C(I) \le \min(S(I_1), S(I_2)) \le \max(S(I_1), S(I_2)) \qquad (4.1)$$

Therefore, the bounds for $C(I)$ and $F(I)$ given by Eq. (3.11) and Eq. (3.12) remains the same. The bounds for $F(I)$ in Eq. (3.17) is then replaced by

$$F(I) = C(I) - S(I) \ge 1 - e^{-\frac{\lambda^2}{2V}(1+o(1))} - 2e^{-\lambda_{\min}^2(1+o(1))/2V_{\min}} \qquad (4.2)$$

where $\lambda_{\min}$ and $V_{\min}$ correspond to the set I_j whose $S(I_j) = \min(S(I_1), S(I_2))$. In short, we can conclude that relaxing the candidacy requirement doesn't worsen the bound for the success probability. The bound for the failure probability is worsen only when the set I is likely to be infrequent since the relaxed candidacy requirement leads to a worse bound for the candidacy probability and since the bound for the success probability stays the same, the bound for the failure probability worsen.

5 Maximum frequent itemsets: some examples

We will now attempt to use the insights from Section 3 regarding the success probability $S(I)$ to analyze the size of the largest frequent set I under several probability assignments to $P(I)$, the probability that a transaction T contains the set of items I. Our probability model is as follow

1. The items in I are picked at random from the set of items $\mathcal{I}$, with the probability that an item $i \in \mathcal{I}$ belongs to I with probability p.

2. The set $I \subseteq \mathcal{I}$ are such that

$$\Pr(I = \{i_1, i_2, \dots, i_m\}) = \frac{e^{-\mu}\mu^m}{m!\binom{n}{m}}$$

where $n = |\mathcal{I}|$ is the total number of items in $\mathcal{I}$. This corresponds to the situation where the number of items in a transaction T follows a Poisson distribution with mean μ for some constant μ.

In case 1 above, the probability that a transaction T contains the set I is p^m where $m = |I|$. This can be seen easily since

$$\Pr(T \supseteq I) = \Pr(i_1 \in T \cap i_2 \in T \cap \cdots \cap i_m \in T) = \prod_{1 \le k \le m} \Pr(i_k \in T) = p^m$$

Therefore, since $bP(I) < \sigma$ implies that $S(I) \to 0$ as $\sigma - bP(I)$ increases, we have that, in order for I to be frequent,

$$bP(I) = b\Pr(T \supseteq I) = bp^m \ge \sigma \Rightarrow m \le \frac{\log \frac{\sigma}{b}}{\log p} = \theta_1 \tag{5.1}$$

and so for case 1, the size of the largest frequent set is tightly concentrated around θ_1. In fact, if I is a set of size θ_1, then $P(I) = p^{\theta_1}$ and so the variance $V = bp^{\theta_1}(1-p^{\theta_1})$. Since the width of the concentration interval is $O(\sqrt{bp^{\theta_1}(1-p^{\theta_1})} = O(\sqrt{\sigma(b-\sigma)/b})$, if $\sigma = b^{\epsilon}$ for some $\epsilon > 0$, then the concentration interval will have width $b^{\epsilon/2}$.

Case 2 is a little more complicated. If I is a set of items in $\mathcal{I}$ with $|I| = m$, then the set

$$J_{m+k} = \{J \subseteq \mathcal{I} : J \supseteq I \wedge |J| = m + k\} \tag{5.2}$$

has size $|J_{m+k}| = \binom{n-m}{k}$ since given a $J \in J_{m+k}$, we can chose the elements of $J \setminus I$ in $\binom{n-m}{k}$ ways. Since $T \supseteq I$ implies that $T \in J_{m+k}$ for some $k \ge 0$ and that

$$\Pr(J_{m+k}) = \frac{e^{-\mu}\mu^{m+k}\binom{n-m}{k}}{(m+k)!\binom{n}{m+k}} \tag{5.3}$$

we have

$$\begin{aligned}
\Pr(T \supseteq I) &= \sum_{k=0}^{n-m} J_{m+k} \\
&= \sum_{k=0}^{n-m} \frac{e^{-\mu}\mu^{k+m}\binom{n-m}{k}}{(m+k)!\binom{n}{m+k}} \\
&= \sum_{k=0}^{n-m} \frac{e^{-\mu}\mu^{k+m}(n-m)!(n-m-k)!(m+k)!}{n!(m+k)!(n-m-k)!k!} \\
&= \frac{e^{-\mu}\mu^{m}(n-m)!}{n!} \sum_{k=0}^{n-m} \frac{\mu^k}{k!} \\
&\le \frac{e^{-\mu}\mu^{m}(n-m)!}{n!} e^{\mu} \\
&\le \frac{\mu^m}{n(n-1)(n-2)\ldots(n-m+1)} = F(m)
\end{aligned} \tag{5.4}$$

We want $bP(I) = b\Pr(T \supseteq I)$ to be at least as large as σ so that I is likely to be frequent. Now, since $\Pr(T \supseteq I) \le F(m)$, if $F(m) \le \sigma/b$, then I is likely to be infrequent. If we let θ_2 be the value of m such that $F(m) \le \sigma/b$, then the size of the largest frequent set I is tightly concentrated around θ_2.

6 Martingale methods

The success and failure probabilities of the Apriori algorithm was analyzed in Section 3 under the probability assumption that for any given subset I of items, the probability that a transaction $T \in \mathcal{T}$ contains I is $P(I)$, independent of any other transaction $T' \in \mathcal{T}$. This probability model is from [6]. An obvious generalization of the above mentioned probability model will then be to remove the assumption of independence. This then is the goal for our analysis in this section.

The key idea of the analysis in Section 3 is the investigation of Eq. (3.2)

$$S(I) = \Pr\left(\sum_{T \in \mathcal{T}} \xi_I(T) \ge \sigma\right)$$

where the $\xi_I(T)$, $T \in \mathcal{T}$ are independent. If we now remove the assumption of independence of the transactions $T \in \mathcal{T}$ then the above sum is a sum of possibly dependent random variables ξ_T. Our analysis of the modified sum will depend on what is known as martingales and their concentration inequalities. Before describing the notion of martingales, we will give a brief discussion of *conditional expectation*.

Definition 6.1: Let X and Y be two random variables. We define the expression $\mathbb{E}[X|Y=y]$ as

$$\mathbb{E}[X|Y=y] = \sum_x x\Pr(X=x|Y=y) \tag{6.1}$$

where the sum is over all x in the range of X. ♣

From the above definition we can easily show the following result

Proposition 6.2: *For any random variables X and Y*

$$\mathbb{E}[X] = \sum_y \Pr(Y=y)E[X|Y=y] \tag{6.2}$$

where the sum is over all values y in the range of Y and all of the expectations exist.

The conditional expectation is then defined in terms of Definition 6.1 as follow

Definition 6.3: Let X and Y be two random variables. The *conditional expectation* of X given Y, $\mathbb{E}[X|Y]$, is the random variable that takes on the value $\mathbb{E}[X|Y=y]$ whenever $Y=y$. ♣

We will also need the following property of conditional expectation.

Proposition 6.4: *Let X, Y, and Z be random variables. We have*

$$\mathbb{E}[X|Z] = \mathbb{E}[\,\mathbb{E}[X|Y\cap Z]\,|\,Z] \tag{6.3}$$

With the above definitions in place, we can now describe the notion of martingales.

Definition 6.5: A sequence of random variables $Z_0, Z_1, \ldots$ is a *martingale* with respect to the sequence $X_0, X_1, \ldots$ if for all $n \geq 0$ the following conditions hold

- Z_n is a function of $X_0, X_1, \ldots, X_n$
- $\mathbb{E}[|Z_n|] \leq \infty$
- $\mathbb{E}[Z_{n+1}|X_0, X_1, \ldots, X_n] = Z_n$ ♣

It's not assumed in the above definition that the sequences $\{X_i\}$ and $\{Z_i\}$ are distinct. In fact, $\{Z_i\}$ can be a martingale with respect to itself. The simplest example of a martingale is the following gambler's situation. Let $\{X_i\}$ be a sequence of fair games, i.e. $\mathbb{E}[X_i]=0$ for all i, and let Z_i be the sum of the X_j, $j \leq i$. Z_i then represent the total winning of the gambler after i games. Now,

$$\mathbb{E}[Z_{i+1}|X_0, \ldots, X_i] = Z_i + \mathbb{E}[X_{i+1}] = Z_i$$

and so the $\{Z_i\}$ is a martingale with respect to $\{X_i\}$.

Given a finite sequence of random variables $\{X_i\}$, $0 \leq i \leq n$ we can constructed a martingale $\{Z_i\}$ with respect to $\{X_i\}$ by the following process. Let Y be a random variable depending on the $\{X_i\}$. Then for any i, define the random variable Z_i as

$$Z_i = \mathbb{E}[Y|X_0, X_1, \ldots, X_i], \quad \text{for } i = 0, 1, \ldots, n \tag{6.4}$$

The $\{Z_i\}$ is then a martingale with respect to the $\{X_i\}$. We can see this as follows

$$\begin{aligned}\mathbb{E}[Z_{i+1}|X_0, X_1, \ldots X_i] &= \mathbb{E}[\,\mathbb{E}[Y|X_0, X_1, \ldots, X_{i+1}]\,|\,X_0, X_1, \ldots, X_i] \\ &= \mathbb{E}[Y|X_0, X_1, \ldots, X_i] \qquad \text{by Proposition 6.4} \\ &= Z_i\end{aligned}$$

The above construction of the $\{Z_i\}$ is known as the Doob's martingale process. The $\{Z_i\}$ is in essence the expected value of Y as we reveal one by one the elements of the sequence $\{X_i\}$. The application of martingales to our analysis is then as follow. We restate the key equation for the success probability

$$S(I) = \Pr\left(\sum_{T \in \mathcal{T}} \xi_I(T) \geq \sigma\right) \tag{6.5}$$

where the ξ_T are not necessarily independent. Each of the ξ_T is either 1 or 0, representing whether $I \subseteq T$ or $I \nsubseteq T$, respectively. Let's now order the ξ_T arbitrarily from 1 to b where $b = |\mathcal{T}|$. Let S_b be the sum of the ξ_i. We define the random variables Z_i as $Z_i = \mathbb{E}[S_b|\xi_0, \xi_1, \ldots, \xi_i]$ with $\xi_0 = 0$. The $\{Z_i\}$ is then a martingale by the Doob's martingale process. The following martingales concentration inequality is important in many applications of martingales [9]

Theorem 6.6: *Let $Z_0, Z_1, \ldots, Z_n$ be a martingale with $a_k \leq |Z_{k+1} - Z_k| \leq b_k$ for each $0 \leq k \leq n-1$, with suitable constants a_k, b_k. Then for any $t \geq 0$*

$$\Pr(|Z_n - Z_0| \geq t) \leq 2e^{-2t^2/\sum(b_k - a_k)^2} \tag{6.6}$$

For example, the case where the ξ_T are independent leads to

$$0 \leq |Z_{k+1} - Z_k| \leq \max(p, 1-p) = \tau$$

Since $Z_0 = \mathbb{E}[S_b] = bP(I)$, and $Z_n = S_b$ itself, we have

$$\Pr(|S_b - bP(I)| \geq t) \leq 2e^{-2t^2/2b\tau^2}$$

The application of Theorem 6.6 requires that we be able to find suitable constants a_k, b_k such that $a_k \leq |Z_{k+1} - Z_k| \leq b_k$. This might not always be possible in the general case since the dependency of the random variables ξ_k could be such that for some index i, if $\xi_i = 0$ then all other ξ_k is also 0, while if $\xi_i = 1$ then all other ξ_k is also 1. In such a case, the value of a_i and b_i will be too large for the bound in Eq. (6.6) to be useful. However, some simplification assumptions can be made that will improve the applicability of Theorem 6.6 to the martingale $\{Z_i\}$. The most obvious assumption is that the random variables ξ_k are at most r dependent for some small constant r, i.e. that we can partition the ξ_k into blocks where the size of each block is at most r and that the blocks are independent from each other. If this is the case, then we rearrange the ξ_k so that any two ξ_i and ξ_j belonging in the same block of size θ must have $|j - i| \leq \theta$. The sum S_b can now be rewritten as $S_b = \zeta_1 + \zeta_2 + \cdots + \zeta_l$ where l is the number of blocks and the ζ_j are the sum of the ξ_k in the same block and so are independent of each other. The martingale $\{Z_k\}$, $1 \leq k \leq l$ with $Z_k = \mathbb{E}[S_b | \zeta_1, \zeta_2, \ldots, \zeta_k]$ then satisfy $|Z_{k+1} - Z_k| \leq r$ since $|\zeta_j| \leq r$ for any j. Under this assumption, with $Z_0 = \mu$ the expected number of transactions containing I

$$\Pr(|Z_l - Z_0| \geq t) = \Pr(|S_b - \mu| \geq t) \leq 2e^{-2t^2/lr^2} \tag{6.7}$$

where l is the number of blocks. If $\sigma > \mu$, then $S(I)$ decreases rapidly towards 0 as $t = \sigma - \mu$ increases. Otherwise, if $\sigma < \mu$, then $S(I)$ increases rapidly towards 1 as $t = \mu - \sigma$ increases.

Another application of the martingale method is the following probability model. We assume that the set of transactions $\mathcal{T}$ is sampled without replacement from the set of all non-empty subsets of $\mathcal{I}$, i.e. the probability model is the hypergeometric distribution. The set $I \subseteq \mathcal{I}$ is frequent if, out of the b sampled transactions at least σ of the them contains I. Now, the number of transactions that contains I is $2^{|\mathcal{I} \setminus I|}$ while the total number of non-empty subsets of $\mathcal{I}$ is $2^{|\mathcal{I}|} - 1$. The expected number of transactions that contains I is then, from the expectation of the hypergeometric distribution, $\frac{b2^{|\mathcal{I} \setminus I|}}{2^{|\mathcal{I}|}-1}$. We now derive the expression for the martingale $\{Z_k\}$. Let $\nu_k = \xi_1 + \xi_2 + \cdots + \xi_k$. If the first k sampled transactions have ν_k transactions that contains I, then the expected number of transactions that will contain I in the remaining $b - k$ transactions to be sampled will be $\frac{(b-k)(2^{|\mathcal{I} \setminus I|} - \nu_k)}{2^{|\mathcal{I}|}-k-1}$. Therefore

$$\begin{aligned} Z_k &= \mathbb{E}[S_b | \xi_1, \xi_2, \ldots, \xi_k] \\ &= \nu_k + \mathbb{E}[\xi_{k+1} + \xi_{k+1} + \cdots + \xi_b | \xi_1, \xi_2, \ldots, \xi_k] \\ &= \nu_k + \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - \nu_k)}{2^{|\mathcal{I}|} - k - 1} \end{aligned}$$

and so

$$Z_{k+1} - Z_k = v_{k+1} - v_k + \frac{(b-k-1)(2^{|\mathcal{I} \setminus I|} - v_{k+1})}{2^{|\mathcal{I}|} - k - 2} - \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 1} \tag{6.8}$$

Now, $v_{k+1} - v_k = \xi_{k+1} \in \{0, 1\}$. If $\xi_{k+1} = 0$, then $v_{k+1} = v_k$ and so

$$|Z_{k+1} - Z_k| = \left| \frac{(b-k-1)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 2} - \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 1} \right| = a_k$$

Otherwise, if $\xi_{k+1} = 1$, then $v_{k+1} = v_k + 1$ and

$$\begin{aligned}
|Z_{k+1} - Z_k| &= \left| 1 + \frac{(b-k-1)(2^{|\mathcal{I} \setminus I|} - v_k - 1)}{2^{|\mathcal{I}|} - k - 2} - \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 1} \right| \\
&= \left| 1 - \frac{b-k-1}{2^{|\mathcal{I}|} - k - 2} + \frac{(b-k-1)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 2} - \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 1} \right| \\
&\le \left| 1 - \frac{b-k-1}{2^{|\mathcal{I}|} - k - 2} \right| + \left| \frac{(b-k-1)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 2} - \frac{(b-k)(2^{|\mathcal{I} \setminus I|} - v_k)}{2^{|\mathcal{I}|} - k - 1} \right| \\
&= b_k
\end{aligned}$$

By Eq. (6.6), we therefore have

$$\Pr(|Z_b - Z_0| \ge t) = \Pr(|S_b - \mu| \ge t) \le 2 \exp\left\{ -2t^2 / \sum \left(\tfrac{2^{|\mathcal{I}|} - b - 1}{2^{|\mathcal{I}|} - k - 2} \right)^2 \right\} \le 2e^{-2t^2/b} \tag{6.9}$$

where $\mu = \frac{b 2^{|\mathcal{I} \setminus I|}}{2^{|\mathcal{I}|} - 1}$. Once again, if $\sigma > \mu$, then $S(I)$ decreases rapidly towards 0, otherwise if $\sigma < \mu$, then $S(I)$ increases rapidly towards 1.

Acknowledgement: We would like to thank Dirk Van Gucht and Paul Purdom for helpful comments and suggestions regarding the results in this paper.

References

[1] Rakesh Agrawal and Ramakrishnan Srikant. Fast algorithms for mining association rules. Technical report, IBM, 1994. 1

[2] Rakesh Agrawal, Heikki Mannila, Ramakrishnan Srikant, Hannu Toivonen, and A. Inerki Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatesky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 307–328. MIT Press, 1996.

[3] Mario Boley. On approximating minimum infrequent and maximum frequent sets. In *Discovery Science*, volume 4755 of *Lecture Notes in Computer Science*, pages 68–77. Springer, 2007. 2

[4] Herman Chernoff. A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Annals of Mathematical Statistics*, 23:493–507, 1952. 4

[5] Fan Chung and Linyuan Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3:79–127, 2006.

[6] Nele Dexters. *An Analysis of Mining Algorithm in Databases and Streams*. PhD thesis, Universiteit Antwerpen, 2007. 1, 2, 3, 4, 5, 6, 7, 11

[7] Nele Dexters, Paul W. Purdom, and Dirk Van Gucht. A probability analysis for candidate-based frequent itemsets algorithms. In *Proceedings of the 21st ACM Symposium on Applied Computing*, pages 541–545, 2006.

[8] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58:13–30, 1963.

[9] Colin McDiarmid. Concentration. In M. Habib, C. McDiarmid, J. Ramirez, and B. Reed, editors, *Probabilisitc Methods for Algorithmic Discrete Mathematics*, pages 195–248. Springer, 1998. 4, 13

[10] Taneli Mielikäinen. Intersecting data to closed sets with constraints. In *Proceedings of the Workshop on Frequent Set Mining Implementations*, 2003.

[11] Michael Mitzenmacher and Eli Upfal. *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University press, 2005.

[12] Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth. Average case performance of the Apriori algorithm. *Siam Journal on Computing*, 33:1223–1260, 2004.

[13] Mohammed J. Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, and Wei Li. New algorithms for fast discovery of association rules. In *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining*, pages 283–286, 1997. 1, 9