A New Bound on Sampling for Frequent Itemsets Mining

Shiyu Ji
shiyu@cs.ucsb.edu

Abstract

In this paper we present a new error bound on sampling algorithms for frequent itemsets mining.

1 Preliminaries

1.1 Frequency of Itemset

In this paper we use the notations and definitions from Riondato and Upfal's pioneering work [3]. Let $\mathcal{I}$ be the set of items $\mathcal{I} = \{I_1, \cdots, I_N\}$ where $N = |\mathcal{I}|$. A transaction τ is a subset of $\mathcal{I}$ (i.e., $\tau \subseteq \mathcal{I}$). An itemset A is a set of items that appear together in a transaction τ, i.e., $A \subseteq \tau$. Clearly any itemset is also a subset of $\mathcal{I}$. Let $\mathcal{D}$ be the set of all the transactions. Denote by $T_{\mathcal{D}}(A)$ the set of all the transactions in $\mathcal{D}$ that contain the itemset A. $T_{\mathcal{D}}(A)$ is also known as the support set of A in $\mathcal{D}$. If $\mathcal{D}$ is a finite set, we can define the frequency of itemset A in $\mathcal{D}$ as the fraction of transactions in D that contain A.

$$f_{\mathcal{D}}(A) = |T_{\mathcal{D}}(A)|/|\mathcal{D}|.$$

Clearly $0 \leq f_{\mathcal{D}}(A) \leq 1$ for any $A \subseteq \mathcal{I}$.

The goal of our sampling algorithm is to approximate $f_{\mathcal{D}}(A)$ given an itemset A as accurately as possible.

1.2 Approximation Algorithms

An (ϵ, δ)-approximation algorithm of the frequencies $f_{\mathcal{D}}(\cdot)$ takes as input all the items $\mathcal{I}$ and outputs a sampled average $f_{\mathcal{S}}(A)$ for each $A \subseteq \mathcal{I}$ such that with probability at least $1 - \delta$,

$$\sup_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq \epsilon.$$

1.3 Risk Bounds

We briefly review some risk bounds in statistical learning theory [1] with the background of frequent itemsets mining.

For each itemset $A \subseteq \mathcal{I}$, define the indicator function $\phi_A : 2^{\mathcal{I}} \to \{0, 1\}$ as follows.

$$\phi_A(X) = \begin{cases} 1 & \text{if } A \subseteq X \\ 0 & \text{otherwise} \end{cases} \quad B \subseteq \mathcal{I}.$$

Clearly, the frequency $f_{\mathcal{D}}(A)$ the *true* average of $\phi_A(X)$ where X goes over all the transactions in $\mathcal{D}$.

$$f_{\mathcal{D}}(A) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \phi_A(\tau).$$

Similarly let $\mathcal{S}$ be the set of the sampled transactions. Then the *sampled* average of $\phi_A(X)$ can be defined as

$$f_{\mathcal{S}}(A) = \frac{1}{|\mathcal{S}|} \sum_{\tau \in \mathcal{S}} \phi_A(\tau).$$

Clearly $f_{\mathcal{S}}(A)$ is the frequency of A appearing in the sampled transactions $\mathcal{S}$.

Assume $|\mathcal{S}| = n$. For each transaction $\tau_i \in \mathcal{S}$, let σ_i be a Rademacher random variable taking value from $\{-1, 1\}$ with uniform probability distribution. The σ_i's are independent. Assuming $\mathcal{I}$ is finite, we define the sample conditional Rademacher average as follows.

$$\mathcal{R}_{\mathcal{S}} = \mathbb{E}_\sigma \left[\max_{A \subseteq \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi_A(\tau_i) \right],$$

where $\mathbb{E}_\sigma$ denotes the expectation taken over all the random variables σ_i's, conditionally on the sample $\mathcal{S}$.

The following theorem tells us that Rademacher average can be used to upper bound the approximation error, even for the worst case.

Theorem 1. (Theorem 3.2, [1]) For any $\delta > 0$, with probability at least $1 - \delta$,

$$\max_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq 2\mathcal{R}_{\mathcal{S}} + \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

If we want to use the upper bound given in Theorem 1 in an approximation, we still need to upper bound the $\mathcal{R}_{\mathcal{S}}$. A classical result is given by Massart [2].

Theorem 2. (Lemma 5.2, [2]) Let $\ell = \max_{A \subseteq \mathcal{I}} [\sum_{i=1}^{n} \phi_A(\tau_i)^2]^{1/2}$ where each $\tau_i \in \mathcal{S}$. Then

$$\mathcal{R}_{\mathcal{S}} \leq \frac{\ell}{n} \sqrt{2 \log N},$$

where $N = |\mathcal{I}|$ and $n = |\mathcal{S}|$.

Hence we have the following stopping condition for an (ϵ, δ)-approximation sampling algorithm.

$$\Delta = \frac{\ell}{n} \sqrt{2 \log N} + \sqrt{\frac{2 \log(2/\delta)}{n}} \leq \epsilon.$$

However for many applications the above bound is not tight enough [3, 4]. In the next section we will first review the state-of-art bound on the worst approximation error, and then propose a new bound which seems tighter.

2 Refining the Upper Bound

The reason why the bound given in the previous section is often not tight enough in practice is that the ℓ defined in Theorem 2 can be quite large. Suppose there is an itemset A that almost always appears in every transaction in $\mathcal{D}$. Then no matter which sample the algorithm chooses, ℓ is roughly $\sqrt{n}$ and thus the upper bound is larger than $\sqrt{2 \log N / N}$, which converges to zero quite slowly as N grows. For $N = 10000$, the bound is still above 0.028 even all the transactions are sampled.

Riondato and Upfal [3] attempted to give a tighter bound of the Rademacher average $\mathcal{R}_{\mathcal{S}}$.

Theorem 3. (Theorem 3, [3]) Let $w : \mathbb{R}^+ \to \mathbb{R}^+$ be the function defined as

$$w(s) = \frac{1}{s} \log \sum_{A \subseteq \mathcal{I}} \exp \left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2} \right).$$

Then $\mathcal{R}_{\mathcal{S}} \leq \min_{s > 0} w(s)$.

Remark. Note that in Theorem 3, the summation in $w(s)$ takes *exactly* $2^{|\mathcal{I}|}$ terms. However in the original version in [3], the summation can take much less than $2^{|\mathcal{I}|}$ terms. We argue this cannot happen. Based on the proof given in [3], one can reach the inequality as follows.

$$\exp(s\mathcal{R}_{\mathcal{S}}) \leq \sum_{A \subseteq \mathcal{I}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2}\right).$$

Note that on the right hand side, each term in the summation is no less than 1. Hence when taking the logarithm on both sides and dividing by s, each of the $2^{|\mathcal{I}|}$ terms cannot be eliminated. Thus the range of the summation cannot be compressed.

Suppose there is a set $\mathcal{V} \subseteq 2^{\mathcal{I}}$, where $2^{\mathcal{I}}$ denotes the power set of $\mathcal{I}$, such that

$$\alpha(s) := \sum_{A \in 2^{\mathcal{I}}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2}\right) \leq \sum_{A \in \mathcal{V}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2}\right) := \beta(s).$$

We take the limits as s approaches 0.

$$2^{|\mathcal{I}|} = \lim_{s \to 0} \alpha(s) \leq \lim_{s \to 0} \beta(s) = |\mathcal{V}|.$$

Hence $\mathcal{V} = 2^{\mathcal{I}}$.

References

[1] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[2] Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

[3] Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

[4] Matteo Riondato and Eli Upfal. Abra: Approximating betweenness centrality in static and dynamic graphs with rademacher averages. *arXiv preprint arXiv:1602.05866*, 2016.