Analysis of Sampling Techniques for Association Rule Mining

Venkatesan T. Chakaravarthy

Vinayaka Pandit

Yogish Sabharwal

IBM India Research Lab, New Delhi
{vechakra, pvinayak, ysabharwal}@in.ibm.com

ABSTRACT

In this paper, we present a comprehensive theoretical analysis of the sampling technique for the association rule mining problem. Most of the previous works have concentrated only on the empirical evaluation of the effectiveness of sampling for the step of finding frequent itemsets. To the best of our knowledge, a theoretical framework to analyze the quality of the solutions obtained by sampling has not been studied. Our contributions are two-fold. First, we present the notions of ϵ-close frequent itemset mining and ϵ-close association rule mining that help assess the quality of the solutions obtained by sampling. Secondly, we show that both the frequent items mining and association rule mining problems can be solved satisfactorily with a sample size that is independent of both the number of transactions size and the number of items. Let θ be the required support, ϵ the closeness parameter, and $1/h$ the desired bound on the probability of failure. We show that the sampling based analysis succeeds in solving both ϵ-close frequent itemset mining and ϵ-close association rule mining with a probability of at least $(1 - 1/h)$ with a sample of size $S = O(\frac{1}{\epsilon^2 \theta}[\Delta + \log \frac{h}{(1-\epsilon)\theta}])$, where Δ is the maximum number of items present in any transaction. Thus, we establish that it is possible to speed up the entire process of association rule mining for massive databases by working with a small sample while retaining any desired degree of accuracy. Our work gives a comprehensive explanation for the well known empirical successes of sampling for association rule mining.

Categories and Subject Descriptors

F.2 [**Analysis of Algorithms**]: Database Management

General Terms

Algorithms, Theory

Keywords

frequent itemset mining, sampling, association rule mining

1. INTRODUCTION

Association rule mining over the *basket data model* was introduced by Agrawal *et al* [1]. It allows businesses to infer useful information on customer purchase patterns, shelving criterion in retail chains, stock trends etc. The basket data essentially consists of a large number of individual records called *transactions* and each transaction is a list of *items* that participated in the transaction. Consider for example, the database of all the transactions that take place in a retail chain. The goal of association rule mining is to discover rules of the type, "whenever a transaction includes a particular set W of items, it is likely to contain a specific item $I \notin W$". In case of a retail chain, such rules can be used to arrange the items on the shelves to increase the sales of closely related items. A formal definition of association rule mining is presented in Section 2. In this paper, we consider sampling techniques for association rule mining in massive databases.

Sampling has been used quite effectively for solving several problems in databases and data mining. For example, statistics collected from a sample of the database is used to generate near-optimal query execution plans. Another example is fast, approximate computations of aggregate functions over massive databases

Informally, the input to association rule mining consists of the collection of transactions and two parameters $\theta \leq 1$, the required *support* and $\gamma \leq 1$, the desired *confidence*. It consists of two steps, namely *frequent itemset mining* in which itemsets with frequency of at least θ are identified, and *association rule mining* in which the association rules of the type $W \Rightarrow I$, with $I \notin W$, are identified. The itemset $W \cup I$ should have a support of at least θ, and of all the transactions containing W, the fraction of the transactions that contain I should be at least γ. Agrawal and Srikant [2] present the *Apriori* algorithm for frequent itemset mining and *FastGenRules* heuristic to generate the association rules. A substantial body of prior work deals with association rule mining. We refer the reader to the survey by Ceglar and Roddick [4] for more details.

In association rule mining, sampling has been used to speed up frequent itemset mining [7, 6, 8]. The main theme in this line of work is to empirically study the effectiveness of sampling at different sample sizes. Toivonen [7] considers the question of the sample size required to ensure that the frequency of a given itemset in the sample is close to its

actual frequency. However, to the best of our knowledge, none of the previous works address the following question: "What should be the size of the sample so that an algorithm working only on the sample reports all itemsets that have a support of θ and does not report any of the itemsets that have a support of less than θ?". In sampling based approaches, usually one identifies candidate frequent itemsets using the sample and then takes a pass over the database to filter out infrequent itemsets among the candidates. In our framework, we do not allow the algorithm to take a second pass over the database and require that the algorithm works by only looking at the sample. Notice that the above property is desirable, particularly when the database is massive.

The probabilistic nature of sampling based techniques does not allow us to solve the above question exactly. However, it is possible to compute approximate solutions. To that end, we develop the notion of ϵ-*close* solutions. In frequent itemsets mining, given the support parameter θ and the closeness parameter ϵ, a solution is ϵ-close if it reports all itemsets that have a support of θ and does not report any itemset that has a support of less than most $(1-\epsilon)\theta$. In this paper, we consider the following question: "what should be the size of sample so that an algorithm that looks only at the sample reports an ϵ-close solution to the frequent itemsets mining".

We study the sample size required to ensure that an algorithm that works only with the sample can produce ϵ-close solution to the frequent itemset mining problem with high probability. Let N be the number of transactions in the database and m be the number of items. Using simple Chernoff bounds [3] and counting arguments, we show that a sample size of $O(\frac{1}{\epsilon^2\theta}(\min\{m, \Delta + \log N\} + \log h))$ is sufficient, where Δ is the maximum number of items present in a transaction and the algorithm succeeds with probability atleast $(1 - 1/h)$. There are two problems with such a sample size. Firstly, m can be quite large in practice. Secondly we would like to eliminate the dependence on $\log N$ which grows asymptotically with the number of transactions. Our main contribution is an improvement in the sample size over the above naive bound. The question we answer in the paper is, "is it possible to report ϵ-close solutions with a sample size independent of m and N?". We show that it is indeed possible. Our sampling algorithm is a natural one: given the closeness parameter ϵ, it reports all itemsets that have a support of $(1-\epsilon/2)\theta$ in the sample. We show that a sample size $S = O(\frac{1}{\epsilon^2\theta}(\Delta + \log \frac{h}{\theta}))$ is sufficient for our algorithm to compute ϵ-close solutions. The main feature of this improved bound on the sample size is that it is independent of the number of items and number of transactions, and depends only the size of the largest transaction. Our analysis consists of two main ideas: (i) bounding the number of itemsets of a given frequency more carefully than before, and (ii) a technique of analyzing the probability of errors in the geometrically varying ranges in the frequency range of $[1/N, (1 - \epsilon)\theta]$.

We also present the notion of ϵ-close solutions to association rule mining part. We present an algorithm and show that a sample size of $S = O(\frac{1}{\epsilon^2\theta}(\Delta + \log \frac{h}{\theta}))$ is sufficient to produce ϵ-close solutions to association rule mining. Thus, our results present a comprehensive analysis of all aspects of sampling for association rule mining.

1.1 Prior Work

A large body of prior work deals with the association rule mining problem. We refer to the survey by Ceglar and Roddick [4] on various algorithms proposed for solving the problem. Below, we present a brief survey of sampling based ideas proposed in prior work.

Association rule mining is usually carried out in two steps: first, finding frequent itemsets and then using these itemsets to identify the association rules. It is well known that the first step of frequent itemset mining dominates the computational and I/O requirements. Most of the prior work on sampling have concentrated on speeding up this phase by running a frequent itemset mining algorithm only on a small sample of the database. Mostly, they correct the errors in the sampling based output by one or two passes over the database.

Perhaps the first work on sampling for association rule mining is that of Mannila *et al.* [6]. But, their work deals with many issues in addition to sampling and hence, their empirical investigation only points to the possible effectiveness of sampling. It does not provide a comprehensive empirical investigation of the effectiveness of sampling for association rule mining. Toivonen [7] focuses solely on the effectiveness of sampling for frequent itemset mining. He gives a bound on the sample size required to ensure that the frequency of a given itemset in the sample is approximately equal to its frequency in the database. Toivonen also presents detailed experimental evaluation of the sampling techniques on synthetic datasets modeling supermarket data based on [2]. These datasets have 100K records and it is shown that sample sizes from 20,000 to 80,000 give very high accuracy. To eliminate errors induced by sampling he suggests a pass over the entire database and eliminating infrequent itemsets identified by the sampling phase.

Zaki *et al.* [8] also carry out an empirical evaluation of sampling for association rule mining. They observe that, for a given itemset, sample sizes as required by Chernoff bounds [3] to achieve a desired degree of accuracy is independent of the size of database. For databases that have fewer than 400K rows, and for some reasonable accuracy requirements, Chernoff bounds based sample sizes may be as large as the entire database. However, they carry out elaborate empirical experiments to demonstrate that the actual accuracy achieved by sampling is much better than the bounds obtained by theoretical analysis. In particular, their experiments suggest that, to get reasonable accuracy, a heuristic of taking samples of size 10% to 25% of the database is required at various support levels. We observe that when the databases are massive (say all seasonal transactions across all Walmart stores [1]), the sizes suggested by the Chernoff bound analysis are much better than those suggested by the heuristics of Zaki *et al.* [8].

Chen et. al [5] propose a sampling based algorithm for association rule mining, without taking any subsequent passes over the database. They devise sub-sampling based heuristics, wherein a small sub-sample is constructed from a given

---

[1]Report at http://walmartstores.com/FactsNews/ Fact-Sheets/Merchandising_Fact_Sheet.pdf mentions that number of weekly transactions at Walmart stores is 176 million.

random sample and association rule mining is carried over the sub-sample. They present a detailed experimental evaluation of the accuracy of the heuristics. While their algorithms work only on the sample and solve the association rule mining problem, they do not provide any theoretical guarantees on the sample size or accuracy.

Thus, a body of prior work deals with sampling based techniques for the association rule mining problem. However, they do not address the issue of deriving a bound on the sample size required to solve the problem by only looking at the sample, while providing provable guarantees on the accuracy obtained.

2. ASSOCIATION RULE MINING

The input to association rule mining consists of a database T of N transactions, $T = t_1, t_2, \ldots t_N$ over a set of m items $\mathcal{I} = \{I_1, I_2, \ldots I_m\}$. Each transaction is a subset of $\mathcal{I}$. Let Δ denote the size of the largest transaction[2]. A subset $X \subseteq I$ is called an *itemset*. The *frequency* of an itemset is the ratio of the number of transactions that contain the itemset to the total number of transactions in the database. We say that an itemset X is *p-frequent* if its frequency is at least p. The goal of association rule mining is to discover *association rules* of the type $W \Rightarrow I$ where $W \subseteq \mathcal{I}$ and $I \in \mathcal{I} \setminus W$. The *support* of a rule $W \Rightarrow I$ is equal to the frequency of the itemset $W \cup I$. The *confidence* of a rule $W \Rightarrow I$ is equal to the ratio of the frequency of $W \cup I$ to the frequency of W. Given two parameters, $\theta \leq 1$ (called as *support threshold*) and $\gamma \leq 1$ (called as *confidence threshold*), the goal of association rule mining is to discover association rules that have a support of at least θ and a confidence of at least γ. The intuitive meaning of such a rule is that transactions that contain $W \cup I$ occurs frequently and a large fraction of the transactions that contain W also contain I.

3. PROBLEM FORMULATION

Typically, the task of association rule mining is carried out in two steps. The first step consists of finding all θ-frequent itemsets in the database. The second step consists of forming the association rules among the frequent itemsets [2]. It has been observed that the first step of identifying the frequent itemsets is the most computation and I/O intensive. Therefore, we first consider the problem of frequent itemset mining where the goal is to mine all θ-frequent itemsets.

As discussed before, previous work on sampling for association rule mining are mainly empirical studies. From a theoretical point of view, they consider a very restricted notion of accuracy. Our goal is to design an algorithm that only looks at a random sample. Ideally, we would like to report all θ-frequent itemsets and not report any itemset with frequency less than θ. However, the probabilistic nature of sampling does not allow us to provide such a strong guarantee. Therefore, we relax the correctness condition and allow the solution to include some of the less than θ-frequent itemsets whose frequency is very close to θ. We also allow the algorithms to have a small probability of failure.

Given an error parameter $\epsilon > 0$, an *ϵ-close* solution to fre-

---

quent itemset mining is a collection of itemsets that includes all θ-frequent itemsets and does not include any itemset whose frequency is less than $(1 - \epsilon)\theta$ (Refer to Figure 1). Notice that the solution may include some itemsets whose frequency falls in the interval $[(1 - \epsilon)\theta, \theta]$. Then, *$\epsilon$-close frequent items mining* is the problem of finding a solution of the above type.

Given a failure parameter h, our goal is to design an algorithm that looks only at a random sample of size S and outputs an ϵ-close solution with a probability of at least $(1 - 1/h)$. The main issue we consider is the size of the sample required to accomplish this. We observe that none of the previous works on sampling for association rule mining have addressed the problem of ϵ-close frequent items mining.

Similarly, we also consider association rule mining. Given an error parameter ϵ, an ϵ-close solution to association rule mining is one which

- consists of all association rules having support θ and confidence γ

- does not consist of any association rule having support less than $(1 - \epsilon)\theta$

- does not include any association rule having confidence less than $(1 - \epsilon)\gamma$

The problem of finding such a solution is called *ϵ-close association rule mining*.

Given a failure parameter h, our goal is to design an algorithm that looks only at a random sample of size S and outputs an ϵ-close solution to the association rule mining with a probability of at least $(1 - 1/h)$. Again, we consider the bound on the size of the sample required to accomplish this. None of the previous have even considered sampling for the confidence part of association rule mining.

Remark. In this paper, all our samples are obtained with replacement. This is to ensure the independence of random variables while applying Chernoff bounds.

3.1 ϵ-close Frequent Itemset Mining: a Simple Bound

In this section, we derive a simple bound on the sample size that is sufficient for solving the ϵ-close frequent itemset mining. Our algorithm is as follows: Given a sample of size S, it reports all itemsets that occur at least $S(1 - \epsilon/2)\theta$ times in the sample.

We now present the two Chernoff bounds that we use extensively in this paper. Let X be a random variable obtained by summing a set of independent identical indicator variables. Following Chernoff bounds on the deviation of X from its expected value are well known [2]. For all $0 < \delta < 1$:

$$\mathbf{Pr}[X \leq (1 - \delta)\mathbb{E}[X]] \leq e^{-\frac{\delta^2 \mathbb{E}[X]}{2}} \qquad (1)$$

and

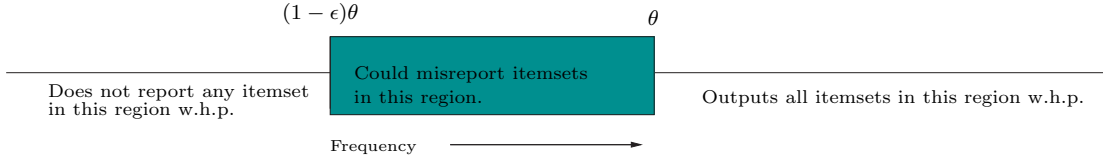$$\mathbf{Pr}[X \geq (1 + \delta)\mathbb{E}[X]] \leq e^{-\frac{\delta^2 \mathbb{E}[X]}{3}} \qquad (2)$$

$(1-\epsilon)\theta$ θ

Could misreport itemsets in this region.

Does not report any itemset in this region w.h.p.

Outputs all itemsets in this region w.h.p.

Frequency

Figure 1: Reporting feature of ϵ-close frequent items mining.

For a given itemset W, let f_W be its frequency in the database and f_W^s be its frequency in the sample. Let $C_W^s = f_W^s \cdot S$ denote the count of W in the sample. Note that $\mathbb{E}[C_W^s] = f_W \cdot S$. It is easy to see that,

$$\mathbf{Pr}[f_W^s \le (1-\delta)f_W] = \mathbf{Pr}[C_W^s \le (1-\delta)\mathbb{E}[C_W^s]]$$

Using the Chernoff bound in Equation 1, for all $0 < \delta < 1$,

$$\mathbf{Pr}[f_W^s \le (1-\delta)f_W] \le e^{-\frac{\delta^2 S f_W}{2}}. \tag{3}$$

Similarly, for all $0 < \delta < 1$,

$$\mathbf{Pr}[f_W^s \ge (1+\delta)f_W] \le e^{-\frac{\delta^2 S f_W}{3}} \tag{4}$$

Note that a θ-frequent itemset W is not reported by our algorithm only if $f_W^s < (1-\epsilon/2)\theta$. Therefore, by invoking Equation 3 with $\delta = \epsilon/2$, we get that,

$$\mathbf{Pr}[W \text{ is not reported}] \le e^{-\frac{\epsilon^2 S \theta}{8}},$$

Similarly, an itemset W with frequency less than $(1-\epsilon)\theta$ is reported by our algorithm only if $f_W^s \ge (1-\epsilon/2)\theta$. Therefore, by invoking Equation 4 with $\delta = \epsilon/2$, we get that,

$$\mathbf{Pr}[W \text{ is reported}] \le e^{-\frac{\epsilon^2 S \theta}{12}}.$$

Clearly, there are at most 2^m itemsets possible where m is the number of items. On the other hand, observe that the number of itemsets that can occur in the database is at most $N2^\Delta$, where Δ is the size of the largest transaction and N is the number of transactions in the database. We observe that a maximum of either 2^m or $N2^\Delta$ itemsets are there in the database. Now, applying the union bound, we get that

$$\mathbf{Pr}[\text{the solution is not } \epsilon\text{-close}] \le \min\{2^m, N2^\Delta\}e^{-\frac{\epsilon^2 S \theta}{12}}.$$

Thus, if we take $S \ge \frac{12}{\epsilon^2 \theta}\min\{m + \log h, \log N + \log h + \Delta\}$, we get that the probability of failure is at most $1/h$.

3.2 Main Question

Note that, for a single set W, the required sample size (given by $\frac{2\log h}{\epsilon^2 \theta}$) is independent of the database size. However, a straightforward application of Chernoff bounds with union bound as shown previously yields a sample size that depends on m or $\log N$. In practice, m can be a few thousands. Moreover, $\log N$ is dependent on the database size and is unsatisfactory in an asymptotic sense. The main question we address in this paper is whether it is possible to solve ϵ-close frequent items mining and ϵ-close association rule mining with a sample size independent of m and N.

Our main contribution is to show that both ϵ-close frequent items mining and ϵ-close association rule mining can be solved with samples whose sizes are independent of m and N. We show that $S \ge O(\frac{1}{\epsilon^2 \theta}(\Delta + \log h/\theta))$ is sufficient to solve ϵ-close frequent items mining and ϵ-close association rule mining. Our results constitute a comprehensive analysis of sampling techniques for association rule mining. The claims are formalized below.

In Section 4, we prove the following theorem about the sample size required to solve ϵ-close frequent itemset mining.

THEOREM 3.1. *There exists an algorithm that, given a sample of size*

$$S \ge \frac{24}{\epsilon^2(1-\epsilon)\theta}\left[\Delta + 5 + \log\frac{5h}{(1-\epsilon)\theta}\right],$$

solves the ϵ-close frequent mining problem, i.e., outputs all θ-frequent itemsets and does not output any of the less than $(1-\epsilon)\theta$-frequent itemsets, with a success probability of at least $(1-\frac{4}{5h})$.

In Section 5, we prove the following theorem about the sample size required to solve ϵ-close association rule mining.

THEOREM 3.2. *There exists an algorithm that, given a sample of size*

$$S \ge \frac{48}{\epsilon^2(1-\epsilon)\theta}\left[\Delta + 5 + \log\frac{5h}{(1-\epsilon)\theta}\right],$$

solves both ϵ-close frequent itemset mining and ϵ-close association rule mining with a success probability of at least $(1-1/h)$.

4. SAMPLING SCHEME AND ANALYSIS
In this section, we present an analysis of sampling for solving ϵ-close frequent items mining. We first present the algorithm used to report the frequent itemsets from the given sample.

4.1 The Algorithm
Given a tolerance range of $0 < \epsilon < 1$, we solve the ϵ-close frequent item mining as follows. Let $\alpha = \epsilon/2$. Our algorithm is same as the one presented in Section 3.1. That is,

- Pick a random sample of size S (with replacement) from the database.

- Report every itemset in S that has a frequency of $(1-\alpha)$, i.e., occurs $S(1-\alpha)\theta$ times in the sample.

We note that the main obstacle for obtaining a sample size independent of m and $\log N$ in the previous section was the

liberal counting of the number of frequent itemsets. It overlooked the fact the number of itemsets of a frequency x is inversely proportional to x. To be able to use this observation for our purpose, we also need to analyze the non-frequent itemsets more carefully than before. We highlight the main ideas of the analysis as we encounter them.

4.2 Accepting all Frequent Itemsets

We now present an analysis that bounds the size of the sample required to ensure that our algorithm reports all θ-frequent itemsets with a high probability. Our analysis is similar to the one presented in Section 3.1 except that we bound the number of θ-frequent itemsets more carefully.

As before, let f_W denote the frequency of a itemset W in the database and f_W^s denote its frequency in the sample. As seen earlier, for a θ-frequent itemset, we have that

$$\mathbf{Pr}[f_W^s < (1 - \alpha)\theta] \leq e^{-\alpha^2 S\theta/2}.$$

Let Failure$_{\geq \theta}$ denote the event that some θ-frequent itemset is not reported by the algorithm.

$$\text{Prob}[\text{Failure}_{\geq \theta}] \leq (\# \, \theta\text{-frequent itemsets})(e^{-\alpha^2 S\theta/2}).$$

We want S to be such that Prob[Failure$_{\geq \theta}$] $\leq \frac{1}{5h}$, ie,

$$(\# \, \theta\text{-frequent itemsets})(e^{-\alpha^2 S\theta/2}) \leq \frac{1}{5h}. \qquad (5)$$

Now, if we use the bound of 2^m or $N2^\Delta$ for the number of θ-frequent itemsets, then we get the same result as in previous section. Here, we use a simple, but what turns out to be powerful observation (especially in analyzing non-frequent itemsets) about the number of itemsets of a given frequency.

LEMMA 4.1. *For any $\beta > 0$, the number of β-frequent itemsets is at most $2^\Delta/\beta$.*

PROOF. There are at most $N2^\Delta$ itemsets in all (with repetition) over all the transactions. An β frequent itemset has to consume at least $N\beta$ of these sets. Therefore there can be at most $(N2^\Delta)/(N\beta) = 2^\Delta/\beta$ of β-frequent itemsets. $\square$

Now, substituting the number of θ-frequent itemsets in Equation 5 by the bound implied by Lemma 4.1 for $\beta = \theta$, we get that S has to be such that,

$$(2^\Delta/\theta)(e^{-\alpha^2 S\theta/2}) \leq \frac{1}{5h}.$$

Solving for S, we get that a sample size of $S = \frac{2}{\alpha^2 \theta}(\Delta + \log \frac{5h}{\theta})$ is sufficient to guarantee the required probability of success. Thus, we have the following lemma,

LEMMA 4.2. *If the sample size $S \geq \frac{2}{\alpha^2 \theta}(\Delta + \log \frac{5h}{\theta})$, Pr[Failure$_{\geq \theta}$] $\leq 1/5h$.*

4.3 Rejecting Non-Frequent Itemsets

We now turn our attention to the itemsets whose frequency is below $(1-\epsilon)\theta$. We want to bound the sample size required

to ensure that, with high probability, the algorithm does not report any itemset with frequency less than $(1-\epsilon)\theta$. Our aim is to prove a bound on the sample size that is independent of m and N.

A straight-forward approach is as follows. First, upper bound the frequency of any itemset in the range $[1/N, (1 - \epsilon)\theta]$ by $(1-\epsilon)\theta$. Second, bound the number of such itemsets by $N2^\Delta$ or 2^m. Third, apply the Chernoff bound along with the union bound to obtain a bound on the sample size. This was essentially the approach used in Section 3.1 and fails to give a bound independent of m and N.

Consider an itemset with frequency $f < (1 - \epsilon)\theta$ in the original database. It gets reported by our algorithm if its frequency in the sample crosses the threshold frequency of $(1-\alpha)\theta$. As f decreases, the following two effects take place: (i) the probability that our algorithm reports such an itemset decreases, and (ii) the bound on the number of itemsets of frequency atmost f, as given by Lemma 4.1 increases. The main observation that we exploit in our analysis is that the rate of decrease in probability is much more than the rate of increase in the number of itemsets in the $(1/N, f)$ range. We therefore split the range of $[1/N, (1 - \epsilon)\theta]$ into multiple geometric ranges and bound the probability of failure separately for each range. The main details are below.

Let $\phi = (1 - \epsilon)\theta = (1 - 2\alpha)\theta$. We divide the range $(1/N, \phi)$ into sub-ranges $R_0, R_1, \ldots, R_{L-1}$ where $L = \log N\phi$. Let R_j denote the sub-range $[\phi/2^{j+1}, \phi/2^j]$ (Refer to Figure 2). For each j, we consider the itemsets in the frequency range R_j and analyze the probability that some itemset in that range is reported by the algorithm.

Consider a range $R_j, j \geq 3$. Let W be an itemset whose frequency belongs to the R_j. Let f_W^s denote its frequency in the sample and f_W denote its frequency in the database. The itemset W is reported by our algorithm if its frequency in the sample $f_W^s \geq (1 - \epsilon/2)\theta$. Recall that $(1 - \epsilon/2)\theta = (1 - \alpha)\theta$ and $\phi = (1 - 2\alpha)\theta$. Therefore, it follows that $f_W^s \geq (1 + \alpha)\phi$. Note that $\phi/2^{j+1} < f_W \leq \phi/2^j$. Therefore, the probability that W is reported is given by

$$
\begin{aligned}
\mathbf{Pr}[f_W^s \geq (1 - \epsilon/2)\theta)] &\leq &\mathbf{Pr}[f_W^s \geq (1 + \alpha)\phi] \\
&\leq &\mathbf{Pr}[f_W^s \geq 2^j(1 + \alpha)f_W] \\
&\leq &\mathbf{Pr}[f_W^s \geq 2^j(1 + \alpha)\mathbb{E}[f_W^s]].
\end{aligned}
$$

Now we use the general form of Chernoff bound [3]: For a random variable X obtained by summing independent indicator random variables, $\mathbf{Pr}[X \geq (1 + \delta)\mu] \leq \left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu$ where $\mu = \mathbb{E}[X]$. Here, $\delta = 2^j(1+\alpha) - 1$ and $\mu \geq S\phi 2^{-(j+1)}$. Therefore,

$$
\begin{aligned}
\mathbf{Pr}[W \text{ is reported}] &\leq &\left(\frac{e^\delta}{(1+\delta)^{(1+\delta)}}\right)^\mu \\
&\leq &\left(\frac{e}{1+\delta}\right)^{(1+\delta)\mu} \\
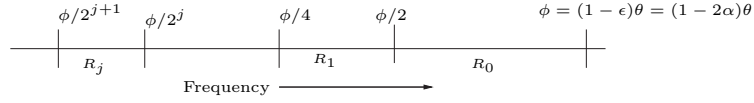&\leq &\left(\frac{e}{2^j(1+\alpha)}\right)^{2^j(1+\alpha)S\phi 2^{-(j+1)}}
\end{aligned}
$$

Figure 2: Geometric ranges over which the non-frequent itemsets are analyzed.

$$\leq \left(\frac{1}{2^{j-2}(1+\alpha)} \right)^{(1+\alpha)S\phi/2}$$

$$\leq \left(\frac{1}{2^{j-2}} \right)^{(1+\alpha)S\phi/2}$$

$$\leq 2^{-(j-2)\phi S/2}$$

Therefore, if we take a sample size $S \geq \frac{2}{\phi}(\Delta + \log \frac{5h}{\phi} + 5)$, then

$$\mathbf{Pr}[W \text{ is reported}] \leq 2^{-(j-2)(\Delta + \log \frac{5h}{\phi} + 5)} \qquad (6)$$

We now prove the following lemma about the sample size required to ensure that, with high probability, no itemset with frequency in the range R_j is reported by our algorithm.

LEMMA 4.3. *Consider the Interval $R_j = (\phi/2^{j+1}, \phi/2^j]$ for $j \geq 3$. If the sample size $S \geq \frac{2}{\phi}(\Delta + \log \frac{5h}{\phi} + 5)$, then the probability that the algorithm reports any itemset having frequency in the range R_j is at most $2^{-(j-2)} \cdot \frac{1}{5h}$, where $\alpha = \epsilon/2$.*

PROOF. From Lemma 4.1, the total number of sets that can occur in R_j in the database is at most $2^{\Delta+j+1}/\phi$. We now use the bound on the probability of failure for a single itemset given by Equation 6 and apply union bound over the $2^{\Delta+j+1}/\phi$ possible itemsets in R_j. The probability that any one of these itemsets is reported by the algorithm, denoted by Prob[Failure$_j$], is

$$
\begin{aligned}
\mathbf{Pr}[\text{Failure}_j] &\leq 2^{-(j-2)(\Delta+\log 5h/\phi+5)} \cdot \frac{2^{\Delta+j+1}}{\phi} \\
&\leq 2^{-(j-3)(\Delta+\log 5h/\phi+5)} \times \\
&\quad 2^{-(\Delta+\log 5h/\phi+5)} \times \frac{2^{\Delta+j+1}}{\phi} \\
&\leq \frac{1}{5h} \cdot 2^{-(j-3)(\Delta+\log 5h/\phi+5)-5+(j+1)} \\
&\leq \frac{1}{5h} \cdot 2^{-7(j-3)-5+(j+1)} \quad \text{as } \Delta, \log \frac{5h}{\phi} \geq 1 \\
&\leq \frac{1}{5h} \cdot 2^{-6j+17} \\
&\leq \frac{1}{5h} \cdot 2^{-(j-2)} \qquad \text{for } j \geq 3
\end{aligned}
$$

This completes the proof of the lemma. $\square$

We now bound the probability of failure for the itemsets in the range $(1/N, \phi/8]$ denoted by Prob[Failure$_{\leq \phi/8}$].

LEMMA 4.4. *Consider the range $R = (1/N, \phi/8]$. If the sample size $S \geq \frac{2}{\phi}(\Delta + \log \frac{5h}{\phi} + 5)$, then the probability that any of the itemsets in the range $(1/N, \phi/8)$ is reported by our algorithm, denoted Prob[Failure$_{\leq \phi/8}$], is at most $\frac{1}{5h}$.*

PROOF. From Lemma 4.3, we have that

$$\mathbf{Pr}[\text{Failure}_{\leq \phi/8}] \leq \sum_{j=3}^{\log N\phi} 2^{-(j-2)} \cdot 1/5h \leq 1/5h.$$

$\square$

Now we consider the remaining itemsets in the range $(\phi/8, \phi]$. We split this range into two ranges $(\phi/8, \phi/2]$ and $(\phi/2, \phi]$. We first consider the range $(\phi/2, \phi]$. Let W be an itemset in the range $(\phi/2, \phi]$. Let f_W^s be its frequency in the sample, and let f_W be its frequency in the database. For W to be reported by our algorithm, its frequency must be at least $f_W^s \geq (1-\alpha)\theta$. Therefore, $\frac{f_W^s}{f_W} \geq 1 + \frac{\alpha}{1-2\alpha}$. The number of itemsets in this range is at most $2^{\Delta+1}/\phi$. Now, following the steps of the proof of Lemma 4.2, we can show that:

LEMMA 4.5. *If the sample size $S \geq \frac{6}{\alpha^2 \phi}(\Delta + 1 + \log \frac{5h}{\phi})$, the probability that our algorithm will report an itemset in the range $(\phi/2, \phi]$, denoted by Prob[Failure$_{(\phi/2,\phi]}$] is at most $\frac{1}{5h}$.*

Compared to the proof of Lemma 4.2, the only change required to prove Lemma 4.5 is the use of Chernoff bound for increase in frequency in the sample. For an itemset W with $f_W \in (\phi/2, \phi]$, we use the following Chernoff bound on the relative increase of the frequency of W in the sample:

$$\mathbf{Pr}[f_W^s \geq (1+\delta)f_W] \leq e^{-\delta^2 S\phi/6} \text{ where } \delta = \alpha/(1-2\alpha).$$

Let us now consider the range $(\phi/8, \phi/2]$. For an itemset W in the range $(\phi/8, \phi/2]$ it is easy to show that $\frac{f_W^s}{f_W} \geq 2 + \frac{2\alpha}{(1-2\alpha)}$ and the number of itemsets in this range is at most $2^{\Delta+3}/\phi$. Using a δ close to 1, it follows that,

LEMMA 4.6. *If the sample size $S \geq \frac{24}{\phi}(\Delta + 3 + \log \frac{5h}{\phi})$, the probability that our algorithm will report an itemset in the range $(\phi/8, \phi/2]$ denoted by Prob[Failure$_{(\phi/8,\phi]}$] is at most $\frac{1}{5h}$.*

4.4 Proof of Theorem 3.1

Now, the algorithm fails if either a θ-frequent itemset is not reported or some itemset that is less than $(1-\epsilon)\theta$-frequent is reported. Observe that the sample size $S = \frac{24}{(\epsilon^2(1-\epsilon)\theta)}\left(\Delta + \log \frac{5h}{((1-\epsilon)\theta)} + 5\right)$ mentioned in the statement of Theorem 3.1 is greater than the samples sizes required in each of the Lemmas 4.2, 4.4, 4.5, and 4.6. Therefore, the

probability of algorithm's failure Prob[Failure] is bounded by

$$
\begin{aligned}
\Pr[\text{Failure}] \;\leq\; & \Pr[\text{Failure}_{\geq \theta}] + \Pr[\text{Failure}_{\leq \phi/8}] \\
& + \Pr[\text{Failure}_{(\phi/8,\phi/2]}] + \Pr[\text{Failure}_{(\phi/2,\phi]}] \\
\leq\; & \frac{1}{5h} + \frac{1}{5h} + \frac{1}{5h} + \frac{1}{5h} \\
\leq\; & \frac{4}{5h}
\end{aligned}
$$

This completes the proof of Theorem 3.1.

5. ASSOCIATION RULE MINING

We now prove Theorem 3.2 for the ϵ-*close association rule mining problem*. Our algorithm for reporting the association rules is simple. It considers only those sets that are output by the ϵ-close frequent items mining. Over them, it employs the standard association rule mining [2] and reports all rules that have a confidence of at least $\frac{(1-\epsilon/4)}{(1+\epsilon/4)}\gamma$ in the sample.

For an association rule $\mathcal{A} \to x$ to have support θ, both the itemsets $\mathcal{A}$ and $\mathcal{A} \cup \{x\}$ must be θ-frequent. Note that our sample size S is larger than the sample size required by the algorithm in Theorem 3.1 and therefore ensures that we identify all the itemsets that have frequency θ and reject itemsets that have frequency less than $(1-\epsilon)\theta$ with a probability of at least $(1-\frac{4}{5h})$. Therefore, we can guarantee that we can reject any association rule not having support at least $(1-\epsilon)\theta$. Now consider the itemsets that are identified as frequent in the previous section. These itemsets have frequency at least $(1-\epsilon)\theta$. For any such itemset I, let f_I denote the frequency of the itemset and $\hat{f}_I$ denote the frequency of the itemset in the sample.

$$
\begin{aligned}
\mathbf{Pr}[|\hat{f}_I - \mathbb{E}[\hat{f}_I]| \geq (\epsilon/4)\mathbb{E}[\hat{f}_I]] \;\leq\; & 2e^{-(\epsilon/4)^2(1-\epsilon)\theta S/3} \\
\leq\; & 2e^{-\epsilon^2(1-\epsilon)\theta S/48}
\end{aligned}
$$

Since there are at most $2^{\Delta}/((1-\epsilon)\theta)$ such itemsets, choosing $S = \frac{48}{\epsilon^2(1-\epsilon)\theta}\left(\Delta + 5 + \log\frac{5h}{(1-\epsilon)\theta}\right)$, we get that $|\hat{f}_I - \mathbb{E}[\hat{f}_I]| \leq (\epsilon/4)\mathbb{E}[\hat{f}_I]$ for all such itemsets with probability at least $(1-\frac{1}{5h})$. Now, our main task is to show that the threshold of $\frac{1-\epsilon/4}{1+\epsilon/4}\gamma$ allows us to distinguish between rules that have confidence of at least γ from those that do not have.

Now consider an association rule $\mathcal{A} \to x$ which has confidence γ. Then $\frac{f_{\mathcal{A}\cup\{x\}}}{f_\mathcal{A}} \geq \gamma$. Therefore,

$$
\begin{aligned}
\frac{\hat{f}_{\mathcal{A}\cup\{x\}}}{\hat{f}_\mathcal{A}} \;\geq\; & \frac{(1-\epsilon/4)\mathbb{E}[\hat{f}_{\mathcal{A}\cup\{x\}}]}{(1+\epsilon/4)\mathbb{E}[\hat{f}_\mathcal{A}]} \\
=\; & \frac{(1-\epsilon/4)}{(1+\epsilon/4)} \cdot \frac{f_{\mathcal{A}\cup\{x\}}}{f_\mathcal{A}} \\
\geq\; & \frac{(1-\epsilon/4)}{(1+\epsilon/4)}\gamma \qquad (7)
\end{aligned}
$$

Next, consider an association rule $\mathcal{A} \to x$ which has confi-

dence $< (1-\epsilon)\gamma$. Then $\frac{f_{\mathcal{A}\cup\{x\}}}{f_\mathcal{A}} < (1-\epsilon)\gamma$. Therefore,

$$
\begin{aligned}
\frac{\hat{f}_{\mathcal{A}\cup\{x\}}}{\hat{f}_\mathcal{A}} \;\leq\; & \frac{(1+\epsilon/4)\mathbb{E}[\hat{f}_{\mathcal{A}\cup\{x\}}]}{(1-\epsilon/4)\mathbb{E}[\hat{f}_\mathcal{A}]} \\
=\; & \frac{(1+\epsilon/4)}{(1-\epsilon/4)} \cdot \frac{f_{\mathcal{A}\cup\{x\}}}{f_\mathcal{A}} \\
<\; & \frac{(1+\epsilon/4)}{(1-\epsilon/4)}(1-\epsilon)\gamma \qquad (8) \\
<\; & \frac{(1-\epsilon/4)}{(1+\epsilon/4)}\gamma
\end{aligned}
$$

The last step follows from the fact that the ratio of $\frac{\hat{f}_{\mathcal{A}\cup\{x\}}}{\hat{f}_\mathcal{A}}$ in Equation 7 is strictly greater than the ratio in Equation in 8 since $(1-\epsilon/4)^2 > (1+\epsilon/4)^2(1-\epsilon)$. Therefore if we report all the association rules having support $(1-\epsilon/2)\theta$ and confidence $\frac{(1-\epsilon/4)}{(1+\epsilon/4)}\gamma$ in the sample S, we satisfy all the required conditions for the ϵ-close association rule mining problem.

This completes the proof of Theorem 3.2.

6. DISCUSSION

As discussed in the introduction, Zaki *et al.* [8] show that, for small databases, a heuristic approach of sampling a fixed percentage of the database (between 10% and 25%) reports 95% of the frequent itemsets for various values of θ. In the case of databases that they consider, this turns out to be more effective than Chernoff bounds based sample sizes. On the other hand, we have shown that both the steps of ϵ-close frequent itemset mining and ϵ-close association rule mining (which put more stringent requirements on the accuracy of the results) can be solved to any desired degree of probability of success with sample sizes that are independent of both the number of items and the number of transactions. In this section, we briefly discuss the implication of our result while doing association rule mining on massive databases.

As an example, let us consider all the transactions in a large retain chain like Walmart in a year. Reports at the Walmart website [3] mention that the number of weekly transactions across all their stores is $176 \cdot 10^6$. The number of all transactions in a year would be to the tune of $9 \cdot 10^9$. As a conservative measure, let us work with a database of $2 \cdot 10^9$ transactions. Note that the sampling heuristic reported in Zaki *et al.* [8] would imply prohibitively large sample sizes. Let us now consider the sample sizes suggested by our analysis. We work with $\theta = 1\%$ (and the bounds get better as θ grows), $\Delta = 20$ (in the typical retail data generated in [2] average transaction size is around 10), and $h = 2^{10}$. If we set $\epsilon = 0.25$, we get sample size to be 3.2×10^6 (which is a small fraction of the database) leading to enormous efficiency gain. Let us say we want to be even more stringent on the region of uncertainty and set ϵ to be 0.1. Even then we get a sample size of just $20 \cdot 10^6$ and obtain highly accurate answers. Our technique has the added advantage of not requiring even a single pass of the database.

---

7. CONCLUSION

We presented a comprehensive theoretical analysis of the sampling technique for the association rule mining problem. We presented the notions of ϵ-close frequent itemset mining and ϵ-close association rule mining. We showed that sampling based technique can solve both the problems using a sample whose size is independent of both the number of items and the number of transactions. From our discussion in Section 6, it follows that an empirical evaluation of the sampling technique on massive databases would be very interesting. The accuracy obtained at the sample sizes suggested by our analysis should be investigated. It would be interesting if similar accuracy can be obtained in practice with much smaller sample sizes.

8. REFERENCES

[1] R. Agrawal, T. Imielinski, and A. Swami. Mining association rules between sets of items in large databases. In *SIGMOD Conference*, pages 207–216, 1993.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *VLDB*, pages 487–499, 1994.

[3] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley, 1992.

[4] A. Ceglar and J. Roddick. Association mining. *ACM Computing Surveys*, 38(2), 2006.

[5] B. Chen, P. Haas, and P. Scheuermann. A new two-phase sampling based algorithm for discovering association rules. In *Proceedings of ACM-SIGKDD Conference on Knowledge Discovery and Data mining (KDD)*, pages 462–468, 2002.

[6] H. Mannila, H. Toivonen, and A. Verkamo. Efficient algorithms for discovering association rules. In *Proceedings of the AAAI workshop on Knowledge Discovery in Databases*, pages 181–192, 1994.

[7] H. Toivonen. Sampling large databases for association rules. In *VLDB*, pages 134–145, 1996.

[8] M. Zaki, S. Parthasarathy, W. Li, and M. Ogihara. Evaluation of sampling for data mining of association rules. In *RIDE*, 1997.