Approximating the Probability of an Itemset being Frequent

Nele Dexters

University of Antwerp, Middelheimlaan 1, 2020 Antwerp, Belgium
nele.dexters@ua.ac.be

Abstract. The frequent itemset mining (FIM) problem is a well-known and important data mining problem, that has been studied in considerable depth during the last few years, mainly experimental. For the analytical study of the average case performance of FIM algorithms, the probability that an itemset is a candidate and the probability that a set is frequent (a success) or infrequent (a failure) are of crucial importance. This paper gives a statistical approach to the success probability for the simple shopping model where every item has the same probability and all the items and all the transactions are independent.

1 Introduction

The frequent itemset mining (FIM) problem, introduced in [1], is a well-known and interesting basic problem at the core of many data mining problems [3]. The problem is, given a large database of basket data, i.e. subsets of a fixed set of items $\mathcal{I}$, and a user-defined support threshold k, to find those sets of items occurring together in at least k baskets. In the last two decades, several different algorithms for solving this problem were proposed and studied experimentically [3].

For the analytical study of the behavior of FIM algorithms, the probability that an itemset is a candidate, a frequent set or an infrequent set is of crucial importance. If the itemset is frequent, it is called a success; if a candidate turns out to be infrequent, it is called a failure. All correct FIM algorithms give the same success probability; it is a property of the data, not of the algorithm. This paper gives a statistical approach to the success probability S_l, the probability that a set consisting of l items is frequent, for the simple model where every item has the same probability p of being chosen, and all the items and all the transactions are independent. Related work [5] computes Chernoff bounds to estimate S_l, but the statistical approach presented in this paper is easier to compute and gives better results. The new approach is based on the approximation of the Binomial Distribution.

In the used shopping model, the probability that a randomly filled basket contains the items $\{1, \ldots, l\}$ is $P(l) = p^l$. The probability that at least k baskets contain l items $\{1, \ldots, l\}$ equals

$$S_l = \sum_{j \geq k} \binom{b}{j} [P(l)]^j [1 - P(l)]^{b-j}.$$

It is the probability that the set $\{1, \ldots, l\}$ is a frequent set.

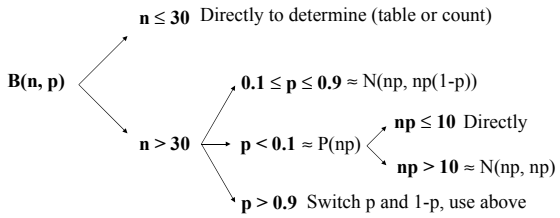

Fig. 1. Overview approximations

2 Statistical Background Information

The Binomial Distribution $X \sim B(n, p)$ is a discrete distribution where X represents the amount of successes in n independent repetitions of a random experiment with two possible outcomes, success with probability p and failure with probability $1 - p$. The probability of having at least k successes in the n successive experiments is

$$P(X \geq k) = \sum_{j \geq k} \binom{n}{j} p^j \ (1-p)^{n-j} = \sum_{j=k}^{n} \binom{n}{j} p^j \ (1-p)^{n-j}.$$

The Binomial Distribution can be approximated by the Normal Distribution and the Poisson Distribution. An overview is given in Figure 1. For more information, see [4] or any other reference book on statistics.

3 Approximation of S_l

In this section, we approximate the success probability S_l, the probability that a set consisting of l items occurs in at least k baskets.

$$S_l = \sum_{j=k}^{b} \binom{b}{j} [P(l)]^j [1 - P(l)]^{b-j}$$

can be seen as $P(X \geq k) = 1 - P(X < k)$ with $X \sim B(b, P(l))$. We now use the appropriate approximation for the Binomial Distribution and investigate the three different situations that can appear. The case $b \leq 30$ is not considered because b is the amount of tuples in the database and this is supposed to be larger than 30.

3.1 $0.1 \leq P(l) \leq 0.9$

We approximate the discrete Binomial distributed $X \sim B\left(b, P(l)\right)$ by the continuous Normal distributed $Y \sim N\left(bP(l), bP(l)(1 - P(l))\right)$. Since we are approximating a

discrete distribution by a continuous one, we have to take care of the continuity correction. The following expression for S_l can be found:

$$S_l \approx 1 - \Phi\left(\frac{(k-0.5) - bP(l)}{\sqrt{bP(l)(1-P(l))}}\right).$$

These values can be found in pre-calculated tables or can be computed using statistical software.

3.2 $P(l) < 0.1$

In this case, the Binomial distributed $X \sim B(b, P(l))$ will be approximated by the Poisson distributed $Y \sim P(bP(l))$. Dependent on de size of $bP(l)$ we can distinguish two different cases.

$bP(l) \leq 10$ In this case, the approximation of the discrete Binomial Distribution by the discrete Poisson Distribution is used. A continuity correction is not necessary. We can find the following expression for S_l:

$$S_l \approx 1 - \sum_{j=0}^{k-1} \frac{(bP(l))^j e^{-bP(l)}}{j!}.$$

$bP(l) > 10$ In this case, the discrete Poisson Distribution itself is approximated by the continuous Normal Distribution and we have to take care of the continuity correction. For S_l, the following expression can be found:

$$S_l \approx 1 - \Phi\left(\frac{(k-0.5) - bP(l)}{\sqrt{bP(l)}}\right).$$

3.3 $P(l) > 0.9$

In this case, $X \sim B(b, P(l))$ with $P(l) > 0.9$. $X' = b - X \sim B(b, 1 - P(l))$ with $1 - P(l) < 0.1$ is constructed. We are now in the previous case (Section 3.2) with X' instead of X. Again, there are two situations that have to be considered.

$b(1 - P(l)) \leq 10$ Analogously as in the previous section, we can find an expression for S_l, based on the approximation by the Poisson Distribution:

$$S_l \approx \sum_{j=0}^{b-k} \frac{(b(1-P(l)))^j e^{-b(1-P(l))}}{j!}.$$

$\boldsymbol{b\,(1 - P(l)) > 10}$ In this case, the Poisson Distribution itself has to be approximated by the Normal Distribution and we find:

$$S_l \approx \Phi\left(\frac{(b-k+0.5) - b(1-P(l))}{\sqrt{b(1-P(l))}}\right).$$

4 Experimental Results

In an experimental study [2], we have compared the approximations with the exact values for S_l for $b = 1024$ and different values of p $(1/2, 1/16)$ and l $(1, 2, 3, 4, 5)$ by means of an error computation. Our study shows that the approximations are accurate when b is large and that they give better results compared to the Chernoff bounds in [5].

During this study of the comparison, we noticed the following interesting facts.

When we look at the exact values for S_l, we can notice that when we fix a certain, moderately-sized value for k, S_l is close to 1 for small values of l and it is close to zero for large values of l. The transition from near 1 to near 0 is quite sharp with increasing l. The transition value of l increases when k decreases. For large k, even S_1 is near zero. For small values of k, l has to be large before S_l approaches zero.

When we look at the approximations for S_l and fix one particular l, the results are more inaccurate when k increases. For increasing l, the approximations are getting worse for smaller and smaller values of k.

In [5], it was found that the values for S with $p = 1/2$ and $l = 4$ are approximately the same as S for $p = 1/16$ and $l = 1$, particularly when k is small. In our approximation, we can see that these values are approximated by the same formula, what in the two cases leads to the same results.

5 Conclusions and Future Work

This paper presents a new, statistically inspired approximation of S_l, the probability that a set consisting of l elements is frequent for the simple shopping model where every item has the same probability and all the items and all the transactions are independent.

Future work includes finding approximations for two other important probabilities: the candidate and the failure probability and applying a new, more realistic and more complex model of shopping behaviour that can cover almost all realistic situations.

References

1. R. Agrawal, T. Imilienski, and A. Swami. Mining association rules between sets of items in large databases. *Proc. ACM SIGMOD, pages 207-216, Washington D.C.*, 1993.
2. N. Dexters. Approximating the probability of an itemset being frequent. *University of Antwerp, Technical Report 2005-03*.
3. B. Goethals. Survey on frequent pattern mining. *www.adrem.ua.ac.be/∼goethals*, 2003.
4. N.L. Johnson and S. Kotz. *Distributions in Statistics: Discrete Distributions*. Houghton Mifflin Company, Boston, 1969.
5. Paul W. Purdom, Dirk Van Gucht, and Dennis P. Groth. Average-case performance of the apriori algorithm. *SIAM J. Computing, vol. 33, No.5, pp. 1223-1260*, 2004.