Fast Approximation of Probabilistic Frequent Closed Itemsets

Erich A. Peterson and Peiyi Tang
University of Arkansas at Little Rock

March 30, 2012

Outline

1. Preliminaries
2. Approximating Probabilistic Frequent Closed Itemsets
3. Experimental Evaluation
4. Conclusions

Preliminaries

Preliminaries

Traditional Itemset Database

T

	a	b	c
t_0	x		x
t_1	x	x	x
t_2		x	
t_3		x	x

- We have a set of items $A = \{a_1, a_2, \ldots, a_m\}$
- We have a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$
- An itemset is any $I \subseteq A$
  - Ex. $I = \{a, b\}$
- Item is either present or not

T

	a	b	c
t_0	x		x
t_1	x	x	x
t_2		x	
t_3		x	x

- The *support* of an itemset I is the number of transactions the itemset occurs in database T, denoted as $Sup_T(I)$
  - Ex. $Sup_T(\{a, c\}) = 2$
- $Sup_{t_j}(I)$ is 1 if $I \subseteq t_j$ or 0 otherwise
- $Sup_T(I) = Sup_{t_0}(I) + Sup_{t_1}(I) + \cdots + Sup_{t_n}(I)$
- Any $I \subseteq A$ whose $Sup_T(I) \geq minsup$ is considered a *frequent* itemset

Uncertain Itemset Database

T

	a	b	c
t_0	0.9		0.21
t_1	0.45	1.0	0.34
t_2		0.88	
t_3		0.6	0.4

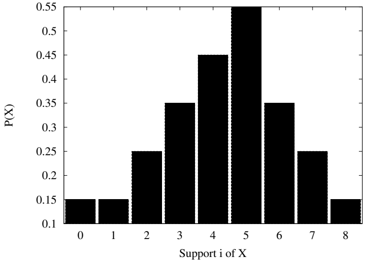
- Each item a has a probability of being in transaction t_j denoted as $Pr(a \in t_j)$
  - Ex. $Pr(a \in t_1) = 0.45$
- $Pr(I \subseteq t_j) = \prod_{a \in I} Pr(a \in t_j)$
  - Ex. $Pr(\{a, b\} \subseteq t_1) = Pr(a \in t_1) \cdot Pr(b \in t_1)$

In Uncertain Databases

- The probability that I occurs in a transaction t_j can be characterized as a Bernoulli random variable X_j^I with parameter $p = Pr(I \subseteq t_j)$
- If $X^I = \sum_{j=0}^{n} X_j^I$, then X^I is a random variable of the Poisson binomial distribution
- $Pr(X^I = i)$ is the probability the support of I is equal i

- Thus, the probability that the support of I is at least $i (X^I \geq i)$ is:

$$Pr(X^I \geq i) = \sum_{k=i}^{n} Pr(X^I = k)$$



- If $Pr(X^I \geq minsup) \geq \tau$, then I is considered a *probabilistic frequent itemset* (PFI) (Bernecker et al.)

Closed Itemset in Traditional Database

- More concise concise / much less redundant output
- If for all itemsets $I' \supset I$, $Sup_T(I') < Sup_T(I)$, then I is closed
- However, there is no concrete support of a uncertain itemset...but we do have the probability

- We defined the new concept of *probabilistic support* (Peiyi Tang et al., ACMSE 2011):

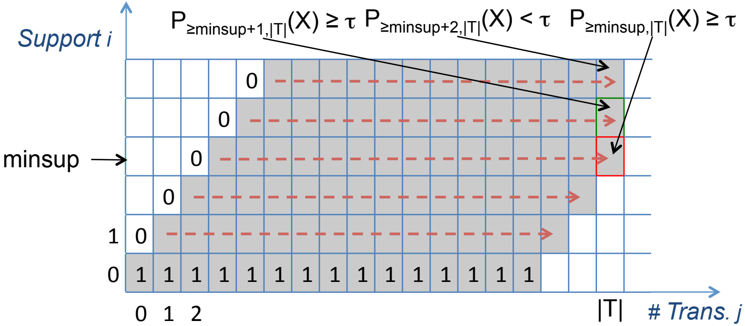
$$PS_T(I, \tau) = argmax_{i \in [0,n]}(Pr(X^I \geq i) \geq \tau)$$

Problem Statement: Probabilistic Frequent Closed Itemset (PFCI)

Given database T and user-defined thresholds τ and *minsup*, mine all itemsets I for which:

- I is probabilistically frequent, i.e. $Pr(X^I \geq minsup) \geq \tau$
- I is closed, i.e. for all $I' \supset I$, $PS_T(I', \tau) < PS_T(I, \tau)$

Each such itemset I we call a *probabilistic frequent closed itemset* (PFCI).



- A dynamic programming approach could be used to calculate $PS_T(I, \tau)$, i.e., with Bernecker et al.'s method
- This can be expensive, as to calculate $PS_T(I, \tau)$ one continues until $Pr(X^I \geq i) < \tau$

Approximating Probabilistic Frequent Closed Itemsets

Approximating Probabilistic Frequent Closed Itemsets

- Wang et al. showed that the Poisson binomial distribution can be approximated using the Poisson distribution
- The Poisson pmf is $Pr(X = i) \approx f(i, \mu) = \frac{\mu^i}{i!} \cdot e^{-\mu}$
  - Thus, the Poisson distribution cdf is $F(i, \mu) = \sum_{k=0}^{i} f(k, \mu)$
- We can use $\mu^I = \sum_{j=1}^{n} \prod_{a \in I} Pr(a \in t_j)$ — the expected support of I in T
- Let $Q(i, \mu^I) = 1 - F(i - 1, \mu^I)$, then $Pr(X^I \geq i) \approx Q(i, \mu^I)$
- $\widehat{PS_T(I, \tau)} = argmax_{i \in [0,n]}(Q(i - 1, \mu^I) \geq \tau)$

Problem Statement: Approx. Probabilistic Frequent Closed Itemset (A-PFCI)

Given an uncertain database T and user-defined threshold τ and *minsup*, mine all itemsets I for which:

- I is an approximate probabilistically frequent itemset, i.e. $\widehat{PS_T(I,\tau)} \geq minsup$
- I is closed, i.e. for all $I' \supset I$, $\widehat{PS_T(I',\tau)} < \widehat{PS_T(I,\tau)}$

Each such itemset is called an *approximate probabilistic frequent closed itemset* (A-PFCI)

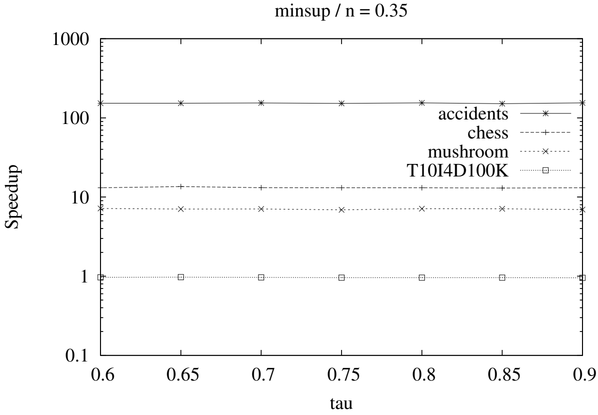
- Let μ_i ($i = 0, \ldots, n$) be the real numbers satisfying $Q(i, \mu_i) = \tau$.
  - i.e. $Q(0, \mu_0) = Q(1, \mu_1) = \cdots = Q(n, \mu_n) = \tau$
  - Because $Q(i, \mu)$ decreases with i and increases with μ: $\mu_0 < \mu_1 < \cdots < \mu_n$
- Using this fact, one can calculate $\widehat{PS(I, \tau)}$ for an itemset I as follows:
- If μ^I satisfies $\mu_i \leq \mu^I < \mu_{i+1}$ for an $i \in [0, n]$, then we have:
  - $\tau = Q(i, \mu_i) \leq Q(i, \mu^I)$
- In addition, we also have the following—for the same reason:
  - $Q(i+1, \mu^I) < Q(i+1, \mu_{i+1}) = \tau$
- This shows that i is the largest value such that $Q(i, \mu^I) \geq \tau$.

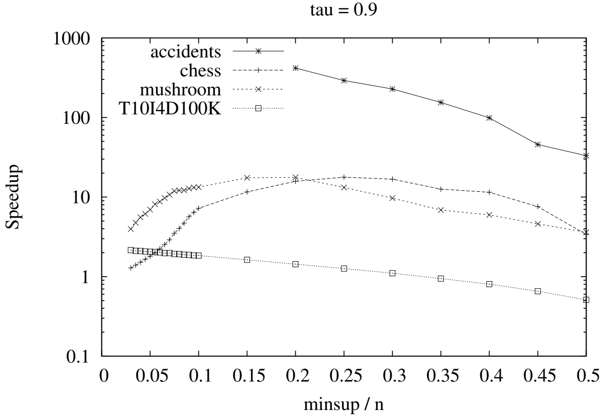
```
function CalcApproxProbSup(itemset I)
    float μ^I ← 0;
    foreach transaction j ∈ T do
        float product ← 1;
        foreach a ∈ I do
            product ← product · T[j][a];
        end foreach
        μ^I ← μ^I + product;
    end foreach
    if μ^I < μ_minsup then
        return −1
    else
        for i = minsup + 1 to n do
            if μ^I < μ_i then
                return i − 1;
            end if
        end for
    end if
end function
```

- Using this method, to calculate $\widehat{PS_T(I,\tau)}$ we need only to "lookup" the right value using the precomputed μ_i $(i = minsup + 1, \ldots, n)$

Experimental Evaluation

Experimental Evaluation


minsup / n = 0.35
1000
100
10
1
0.1
Speedup
0.6
0.65
0.7
0.75
0.8
0.85
0.9
tau
accidents
chess
mushroom
T10I4D100K


tau = 0.9
accidents
chess
mushroom
T10I4D100K
Speedup
minsup / n
1000
100
10
1
0.1
0
0.05
0.1
0.15
0.2
0.25
0.3
0.35
0.4
0.45
0.5

Conclusions

Conclusions

- We define the new concept of an *approximate probabilistic frequent closed itemset* (A-PFCI)
- Will decrease the redundancy and size of output
- Developed an algorithm to mine these new concepts called A-PFCIM

Thank You
Questions?

paper / slides / code
website: erichpeterson.com