Mining Probabilistic Frequent Closed Itemsets in Uncertain Data

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

March 25, 2011

Table of contents

Introduction

Traditional Itemset Database

- Item is either present or not
- Mine those itemsets which are *frequent*—itemsets are frequent if its support is at least *minsup*, $Sup_T(X) \geq minsup$
- Ex. $Sup_T(b, c) = 2$

T

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

Uncertain Itemset Database

- Items have an existential probability of occurring
- There is no support, only probabilities
- How do we calc. support and thus if an itemset is frequent or not?

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

Uncertain Data Model

- Given a set of items $I = \{x_1, x_2, \ldots, x_m\}$
- An itemset is any $X \subseteq I$
- Given a set of transactions $T = \{t_1, t_2, \ldots, t_n\}$
- Each item x has a probability of being in transaction t_j denoted as $P(x \in t_j)$
    - Ex. $P(a \in t_1) = 0.45$
- Possible world semantics are useful in reasoning
    - There exist $2^{|T| \cdot |I|}$ possible worlds for a given database

- If items and transactions are independent the following gives the prob. of a possible world w:

$$P(w) = \prod_{t \in T(w)} (\prod_{x \in t} P(x \in t') \cdot \prod_{x \notin t} (1 - P(x \in t'))) \qquad (1)$$

where $T(w)$ is the set of certain transactions of world w, t a certain transaction in $T(w)$, t' the corresponding *uncertain* transaction in uncertain database T, and $P(x \in t')$ the existential probability of item x in the uncertain transaction t'.

- Thus, we could calc. the probability of itemset X having support i as follows:

$$P_i(X) = \sum_{w \in W, Sup_{T(w)}(X)=i} P(w) \tag{2}$$

where W is the set of possible worlds.

- That would require the enumeration of all possible worlds!

- Bernecker et al. proved you can calc. it as:

$$P_i(X) = \sum_{S \subseteq T, |S|=i} (\prod_{t \in S} P(X \subseteq t) \cdot \prod_{t \in T-S} (1 - P(X \subseteq t))) \quad (3)$$

where T is the original uncertain database and $P(X \subseteq t)$ is

$$P(X \subseteq t) = \prod_{x \in X} P(x \in t)$$

- Thus, the probability of the support of X being at least i is:

$$P_{\geq i}(X) = \sum_{k=i}^{|T|} P_k(X)$$

- Thus, $P_{\geq minsup}(X)$ is the probability that X is frequent
- If this value is above a user-defined confidence threshold τ, then X is considered a *probabilistic frequent itemset*
    - Ex. $P_{\geq minsup}(X) \geq \tau$

Probabilistic Frequent Closed Itemset Mining

What are Closed Itemsets?

- Especially in dense datasets, the number of discovered frequent itemsets can be large
- Mining only maximal itemsets is one solution
  - Is lossy representation: cannot recover all frequent itemset with their support values
- Mining only closed itemsets is another solution
  - Is lossless representation, but there are usually more closed itemsets than maximal itemsets

What is closure / How to calculate it?

- Closure of X Is the largest superset of X that is contained within transactions supporting X
- if for all itemsets $Y \supset X$, $Sup_T(Y) < Sup_T(X)$, then X is closed
- However, there is no concrete supporting transactions...but we do have the probability

T

	a	b	c
t_0	x		x
1_1	x	x	x
t_2		x	
t_3		x	x

- Note that $P_{\geq i}(X) = \sum_{k=i}^{|T|} P_k(X)$ is a non-increasing monotonous function of i, i.e. $P_{\geq j}(X) \leq P_{\geq i}(X)$ for $j > i$.
- We define the new concept of *probabilistic support* as:

$$Sup_T(X, \tau) = argmax_{i \in [0,|T|]}(P_{\geq i}(X) \geq \tau)$$

- Further, $P_{\geq minsup}(X) = \sum_{k=minsup}^{|T|} P_k(X)$ is anti-monotonic. That is, for any $Y \subseteq X$, and any i, $P_{\geq i}(X) \leq P_{\geq i}(Y)$
- Finally, b/c $P_{\geq minsup}(X)$ is anti-monotonic, it can be proven that $Sup_T(X, \tau)$ is as well, i.e. $Sup_T(X, \tau) \leq Sup_T(Y, \tau)$ for any $Y \subseteq X$

- Thus, if an itemset X meets the following two criteria, we consider it a *probabilistic frequent closed itemset* (PFCI):
  1. X is probabilistically frequent, i.e. $Sup_T(X, \tau) \geq minsup$
  2. X is closed, i.e. for all $Y \supset X$, $Sup_T(Y, \tau) < Sup_T(X, \tau)$

Algorithm Sketch

- Bernecker et al. devised a dynamic programming approach for calculating $P_{\geq i,j}(X)$
  - The probability the support of X is at least i in the first j transactions
  - Using the following formula:

$$\begin{aligned} P_{\geq i,j}(X) &= P_{\geq i-1,j-1}(X) \cdot P(X \subseteq t_j) \\ &+ P_{\geq i,j-1}(X) \cdot (1 - P(X \subseteq t_j)) \end{aligned} \tag{4}$$

where $P_{\geq 0,j} = 1 \; \forall . 0 \leq j \leq |T|, P_{\geq i,j} = 0 \; \forall . i > j$

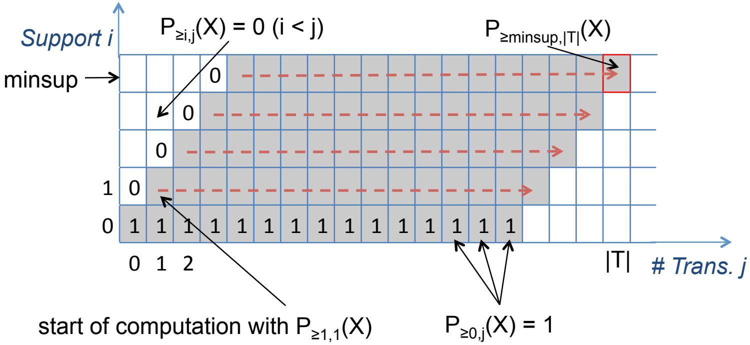


Figure: Calculation of $P_{\geq minsup, |T|}(X)$ Bernecker et al.

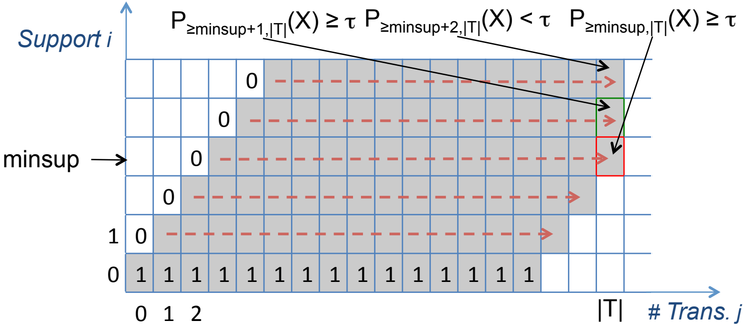
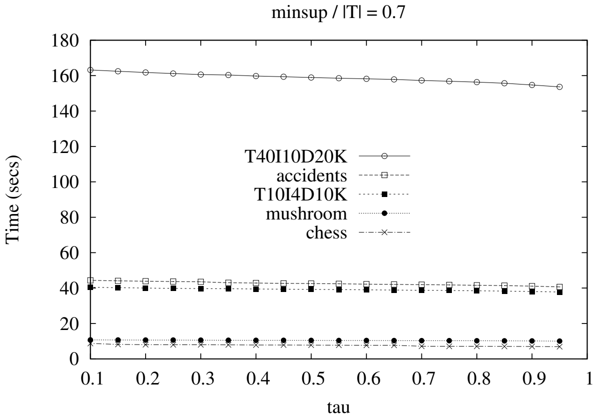
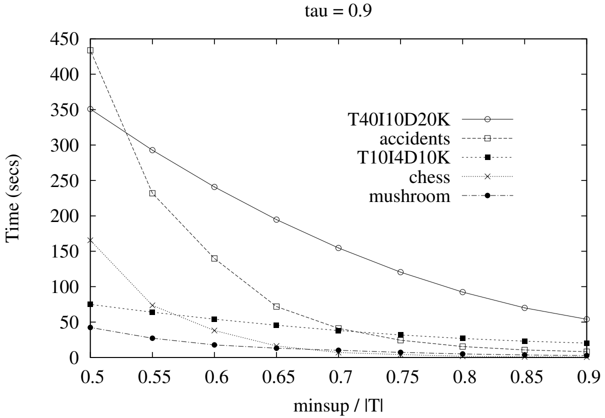


Figure: Calculation of $Sup_T(X, \tau)$

- The previous find if an itemset is probabilistically frequent or not
- The next step in the algorithm, is simply to find out it it is closed as well
- An A-Close like breadth-first method is used—only itemsets of length k and $k+1$ are needed in memory at one time
- Full algorithm details (including pseudocode) can be found in the full paper

Experimental Evaluation




tau = 0.9
Time (secs)
minsup / |T|
T40I10D20K
accidents
T10I4D10K
chess
mushroom
0
50
100
150
200
250
300
350
400
450
0.5
0.55
0.6
0.65
0.7
0.75
0.8
0.85
0.9

Conclusion

- We have introduced a new concept and definition for problem of mining probabilistic frequent closed itemset
- We have introduced an algorithm for mining for these PFCIs
- Currently, we are working on implementing a more efficient algorithm based on the DCI_Closed algorithm

Thank You
Questions?