Mining Probabilistic Generalized Frequent Itemsets in Uncertain Databases

Erich A. Peterson
Peiyi Tang

University of Arkansas at Little Rock

contact@erichpeterson.com

April 4, 2013

Outline

- Uncertain Data Model
- Probabilistic Frequent Itemsets
- Probabilistic Generalized Frequent Itemset Mining[1]
- Conclusion

[1] our contrbution

Uncertain Data Model

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?
- In this research, *possible world semantics* are used

T

TID	Itemset
t_1	$(apple, 0.89), (banana, 0.99), (kale, 1.0)$
t_2	$(apple, 0.4), (banana, 0.45), (cheese, 0.12)$
t_3	$(apple, 0.9), (milk, 0.95), (cheese, 0.20)$

T

TID	Itemset
t_1	$(apple, 0.89), (banana, 0.99), (kale, 1.0)$
t_2	$(apple, 0.4), (banana, 0.45), (cheese, 0.12)$
t_3	$(apple, 0.9), (milk, 0.95), (cheese, 0.20)$

- In possible world semantics, for each uncertain item a in transaction t_j, there exists a concrete instance of t_j which contains a and one that does not
- E.x. Possible World for t_1

$$W(t_1) = \{\langle kale\rangle, \langle apple, kale\rangle, \langle banana, kale\rangle, \langle apple, banana, kale\rangle\}$$

- The probability of an itemset I occurring in an arbitrary transaction t, denoted as $Pr(I \subseteq t)$, is the sum of the probabilities of all the possible worlds $w \in W(t)$ which contain I.
- This requires the enumeration of all possible worlds of $W(t)$; however, $Pr(I \subseteq t)$ can be calculated as:

$$Pr(I \subseteq t) = \prod_{a \in I} Pr(a \in t)$$

T

TID	Itemset
t_1	$(apple, 0.89), (banana, 0.99), (kale, 1.0)$
t_2	$(apple, 0.4), (banana, 0.45), (cheese, 0.12)$
t_3	$(apple, 0.9), (milk, 0.95), (cheese, 0.20)$

E.x. $Pr(\{apple, banana\} \subseteq t_1) = 0.89 * 0.99$

- The set of all possible worlds W induced by *all* transactions in the uncertain database $T\{t_1, \ldots, t_n\}$ is the Cartesian product of $W(t_j)$, $j = 1, \ldots, n$, as follows:

$$W = W(t_1) \times W(t_2) \times \cdots \times W(t_n)$$

tid	itemset
t_1	$(banana, 0.99)$
t_2	$(banana, 0.45), (kale : 0.4)$

	Possible Worlds
$W(t_1)$	$\langle\ \rangle$, $\langle banana \rangle$
$W(t_2)$	$\langle\ \rangle$, $\langle banana \rangle$, $\langle kale \rangle$, $\langle banana, kale \rangle$

TID	Possible Worlds
w_1	$\langle\ \rangle, \langle\ \rangle$
w_2	$\langle\ \rangle$, $\langle banana \rangle$
w_3	$\langle\ \rangle$, $\langle kale \rangle$
w_4	$\langle\ \rangle$, $\langle banana, kale \rangle$
w_5	$\langle banana \rangle$, $\langle\ \rangle$
w_6	$\langle banana \rangle$, $\langle banana \rangle$
w_7	$\langle banana \rangle$, $\langle kale \rangle$
w_8	$\langle banana \rangle$, $\langle banana, kale \rangle$

- If the assumption of independence between the transactions in T is valid, the probability of a possible world $w = (w(t_1), w(t_2), \ldots, w(t_n)) \in W$ can be calculated as follows:

$$Pr(w) = \prod_{i=1}^{n} Pr(w(t_i))$$

where $Pr(w(t))$ was previously defined.

- In an *uncertain* database T, the support of I in transaction t_j, $Sup_{t_j}(I)$, is no longer a concrete 0 or 1. Instead, it is a random variable X_j^I following a Bernoulli distribution with parameter p_j, where $p_j = Pr(X_j^I = 1) = Pr(I \subseteq t_j)$
- The support of I over the entire database T is a random variable $X^I = \sum_{j=1}^{n} X_j^I$
- X^I follows the Poisson binomial distribution with parameters $p_j = Pr(I \subseteq t_j), j = 1, \ldots, n$, if the assumption of independence between transactions is made.
- The probability that $X^I = i$, $(0 \leq i \leq n)$, is:

$$Pr(X^I = i) = \sum_{S \subseteq T, |S| = i} \left(\prod_{t_j \in S} p_j \cdot \prod_{t_j \in T - S} (1 - p_j) \right)$$

Probabilistic Frequent Itemsets

- One can calculate $Pr(X^I \geq minsup)$ using the following formula:

$$Pr(X^I \geq minsup) = \sum_{S \subseteq T, |S| \geq minsup} \left(\prod_{t_j \in S} p_j \cdot \prod_{t \in T - S} (1 - p_j) \right)$$

Definition: Probabilistic Frequent Itemset (PFI)

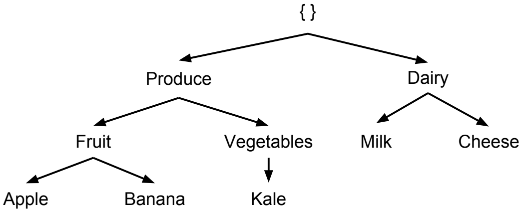
Given an uncertain database T and an itemset I, I is a probabilistic frequent itemset (PFI) with confidence τ, if and only if $Pr(X^I \geq minsup) \geq \tau$, where $minsup \in [0, n]$ and $\tau \in [0, 1]$ are user-defined thresholds. Bernecker et al.

- The problem of mining probabilistic frequent itemsets, is to discover all itemsets I such that $Pr(X^I \geq minsup) \geq \tau$, where $minsup$ and τ are user-defined thresholds.

Probabilistic Generalized Frequent Itemsets

Introduction

- Generalized itemset mining differs from traditional itemset mining, in that the database is accompanied by a taxonomy



- The taxonomy defines the relationships among the items
- With the addition of the taxonomy, new frequent itemsets and association rules may be discovered
    - Apples and Bananas by themselves may not be frequent, but Fruits could be

- Let $D(g) = \{\text{set of descendants of } g \text{ in } G\}$
- The subset and inclusion must be re-defined:

- $a \in_G S$, if and only if:
  - $a \in S$; or
  - $\exists a' \in D(a) : a' \in S$

- $I \subseteq_G S$, if and only if, for each $a \in I$:
  - $a \in_G S$

- E.x. $\{fruit, cheese\}$ is 2, because $\{fruit, cheese\} \subseteq_G t_2$, $\{fruit, cheese\} \subseteq_G t_3$, but $\{fruit, cheese\} \not\subseteq_G t_1$

TID	Itemset
t_1	*apple*, *banana*, *kale*
t_2	*apple*, *banana*, *cheese*
t_3	*apple*, *milk*, *cheese*

Probabilistic Generalized Frequent Itemset Mining

T

TID	Itemset	
t_1	$(apple, 0.89), (banana, 0.99), (kale, 1.0)$	$(fruit, 0.9989)\ldots$
t_2	$(apple, 0.4), (banana, 0.45), (cheese, 0.12)$	$(fruit, 0.995)\ldots$
t_3	$(apple, 0.9), (milk, 0.95), (cheese, 0.20)$	$(fruit, 0.9)\ldots$

- In traditional generalized itemset mining, it is easy to know if the transaction supports a generalized itemset or not
- In an uncertain database, as shown above, we need a way to calculate the probability of a generalized / abstract item occurring in an arbitrary transaction t, i.e., $Pr(g \in_G t)$
- This probability must conform to possible world semantics

- We can formulate a way to calculate $Pr(g \in_G t)$ without the need for enumerating all possible worlds as:

$$Pr(g \in_G t) = 1 - \prod_{a \in D(g)} (1 - Pr(a \in t))$$

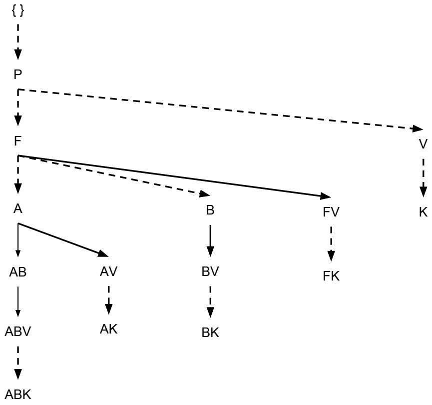
- E.x. Calculating the probability of the generalized item *fruit* occurring in transaction t_1 in out example database, given that $D(fruit) = \{apple, banana\}$, can be done as follows: $Pr(fruit \in_G t_1) = 1 - (1 - Pr(apple \in t_1)) \cdot (1 - Pr(banana \in t_1)) = 0.9989$.

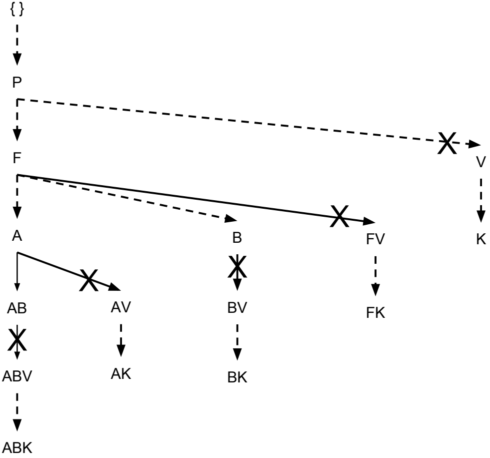
T

TID	Itemset	
t_1	$(apple, 0.89), (banana, 0.99), (kale, 1.0)$	$(fruit, 0.9989) \ldots$

Algorithm Sketch of PGFIM

- Candidate enumeration is done using the SET algorithm (Sriphaew et al.)
  - The relationships between itemsets (subset-superset) and within taxonomy *G* (ascendent-descendent) are used to prune the search space
  - E.x. More generalized items are enumerated first, and the downward closure property is used (smaller subsets are enumerated before larger supersets)

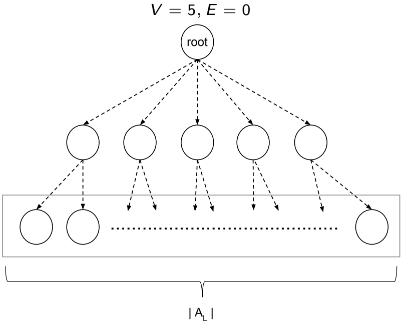




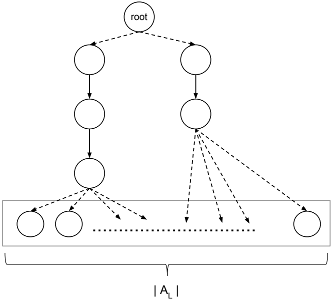
- An exact dynamic programming approach is used to determine frequency (Bernecker et al.)

Taxonomy Generation

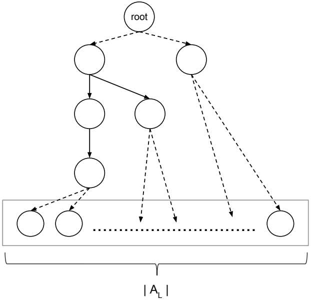
- Diversity of taxonomies experimented with in other's research has tended to be small
- Experimental taxonomies are generated using parameters V and E:
  - V: the number of internal (generalized) vertices found in G
  - E: the number of randomly generated edges connecting the V vertices
- $|A_L|$: the number of attributes (leaf vertices) that have to be present (found in DB)



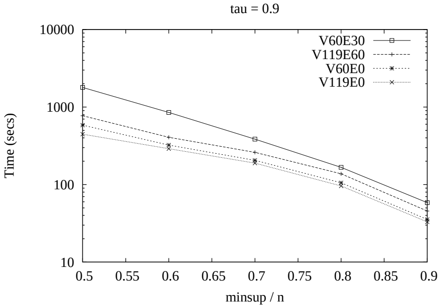
$V = 5, E = 3$

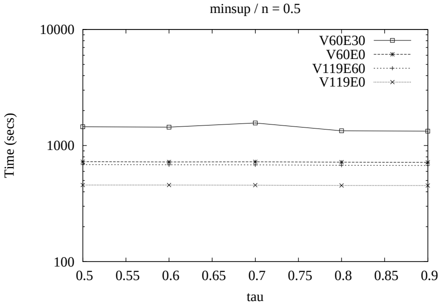


$V = 5, E = 3$

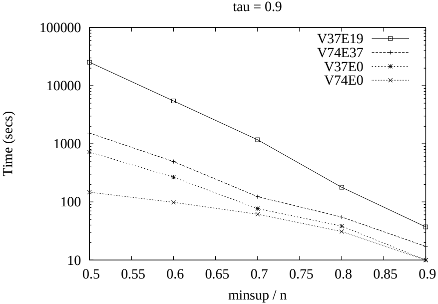


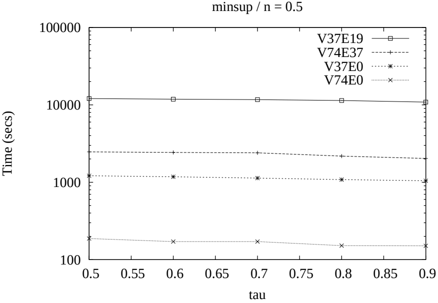
Mushroom Dataset





Chess Dataset





Conclusion

- We have introduced a new concept and definition for the problem of mining probabilistic generalized frequent itemsets (PGFIs)
- An algorithm has been created to mine for such concepts called PGFIM
- Experimental evaluation has been performed

Thank You

Q & A