Discovering Threshold-based Frequent Closed Itemsets over Probabilistic Data

Yongxin Tong[1], Lei Chen[1], Bolin Ding[2]

[1]Department of Computer Science and Engineering, Hong Kong University of Science and Technology

[2] Department of Computer Science, University of Illinois at Urbana-Champaign

Motivations of Probabilistic Frequent Closed Itemsets

TID	Location	Weather	Time	Speed	Probability
T1	HKUST	Rain	8:30-9:00 AM	20-30	0.9
T2	HKUST	Rain	8:30-9:00 AM	null	0.6
T3	HKUST	Rain	8:30-9:00 AM	null	0.7
T4	HKUST	Rain	8:30-9:00 AM	20-30	0.9

TID	Transaction	Probability
T1	a b c d	0.9
T2	a b c	0.6
T3	a b c	0.7
T4	a b c d	0.9

PW	Transactions	Probability	Frequent Closed Itemsets
PW1	T1	0.0108	{ }
PW2	T1, T2	0.0162	{abc}
PW3	T1, T3	0.0252	{abc}
PW4	T1, T4	0.0972	{abcd}
PW5	T1, T2, T3	0.0378	{abc}
PW6	T1, T2, T4	0.1458	{abc} {abcd}
PW7	T1, T3, T4	0.2268	{abc} {abcd}
PW8	T1, T2, T3, T4	0.3402	{abc} {abcd}
PW9	T2	0.0018	{ }
PW10	T2, T3	0.0042	{abc}
PW11	T2, T4	0.0162	{abc}
PW12	T2, T3, T4	0.0378	{abc}
PW13	T3	0.0028	{ }
PW14	T3, T4	0.0252	{abc}
PW15	T4	0.0108	{ }
PW16	{ }	0.0012	{ }

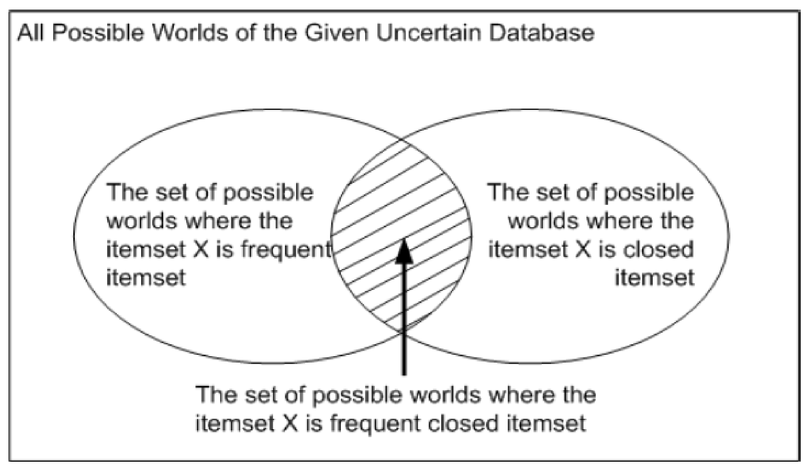
A Uncertain Transaction Database

- Given *min_sup*=2, *pft* = 0.8, there are 15 probabilistic frequent itemsets (7 ones' frequent probability=0.9726, other 8 frequent probabilittis= 0.81).
- How to distinguish 15 itemsets ? **(Probabilistic frequent closed itemset)**
- The frequent closed probability of {abc}=Pr(PW2)+ Pr(PW3)+Pr(PW5)+Pr(PW6) +Pr(PW7)+Pr(PW8)+Pr(PW10)+Pr(PW11)+Pr(PW12)+Pr(PW14)=0.8754.

Definition of Probabilistic Frequent Closed Itemset

- *Frequent Closed Probability*:
  - Given a minimum support *min_sup*, and an itemset X, X's frequent closed probability, denoted as $Pr_{FC}(X)$ is the sum of the probabilities of possible worlds where X is a frequent closed itemset.
- *Probabilistic Frequent Closed Itemset*:
  - Given a minimum support *min_sup*, a probabilistic frequent closed threshold *pfct*, an itemset X, X is a probabilistic frequent closed itemset if:

    Pr{X is frequent closed itemset}= PrFC(X) > pfct
- **It is #P-hard to calculate the frequent closed probability of an itemset when the minimum support is given in an uncertain transaction database.**



The relationship of $Pr_F(X)$, $Pr_C(X)$ and $Pr_{FC}(X)$

MPFCI Algorithm

Procedure MPFCI_Framework {

1: Discover all the single probabilistic frequent items, called Cand, using Chernoff-Hoeffding bound and sort all the items in Cand based on the alphabetic order.

//Phase 1: Construct initial probabilistic frequent single items

2: For each item X in Cand, extend X using a depth-first search like strategy to its supersets with X as prefix and perform Chernoff-Hoeffding bound-based pruning, superset pruning, subset pruning, and frequent closed probability bound-based pruning.

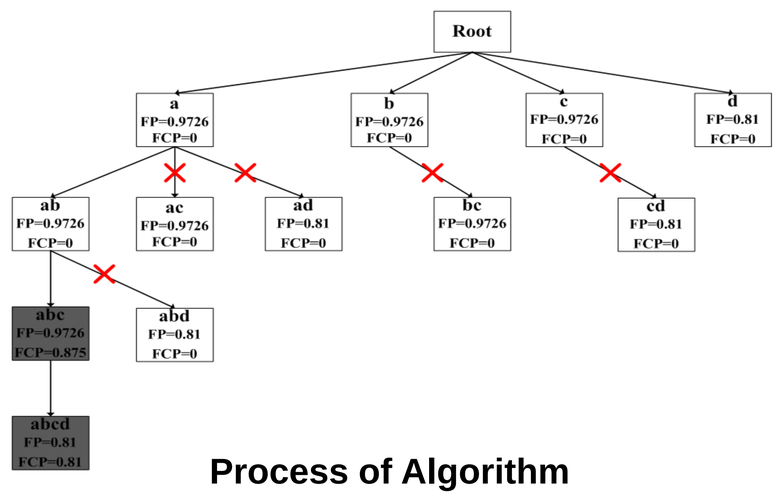
//Phase 2: Bounding and Pruning

3: Check the frequent closed probability of itemsets which cannot be pruned and return the result set.

//Phase 3: Checking

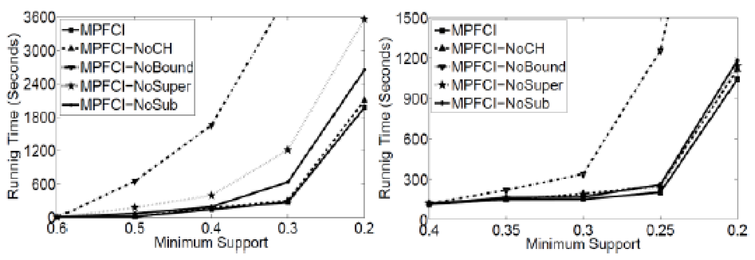
}

- *Four Pruning Methods*:
  - *Chernoff-Hoeffding Bound-based Pruning*
  - *Superset Pruning*
  - *Subset Pruning*
  - *Upper Bound and Lower Bound of Frequent Closed Probability-based Pruning*
- Monte-Carlo sampling algorithm to calculate the frequent closed probability approximately



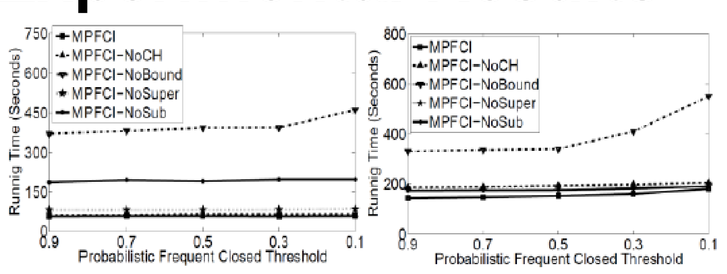
Process of Algorithm

Experimental Results



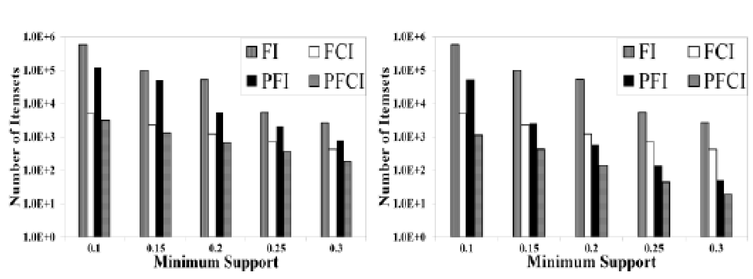
(a) *Mushroom* (b) *T20I10D30KP40*

Running Time w.r.t min_sup



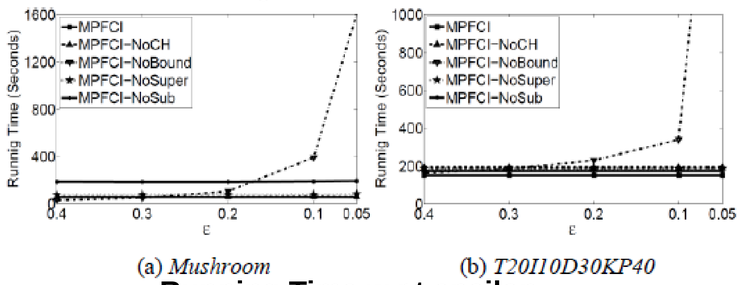
(a) *Mushroom* (b) *T20I10D30KP40*

Running Time w.r.t pfct



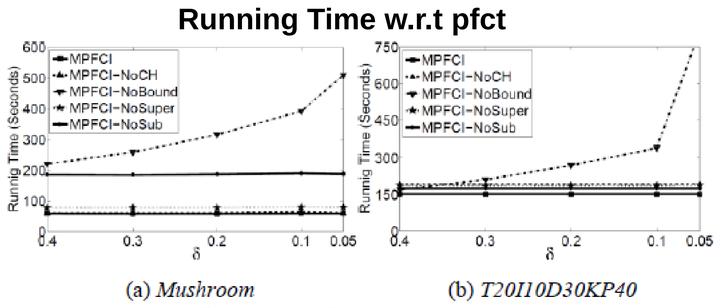
(a) Mean=0.8, Var=0.1 (b) Mean=0.5, Var=0.25

Compression Quality w.r.t min_sup



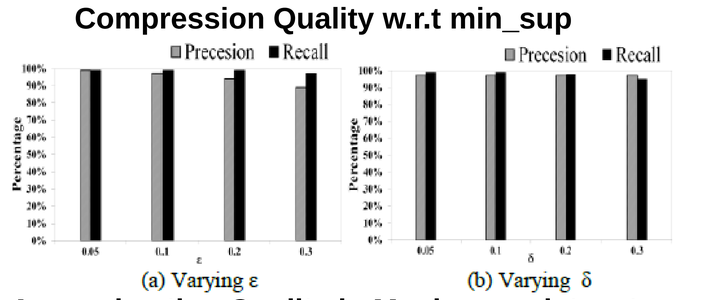
(a) *Mushroom* (b) *T20I10D30KP40*

Running Time w.r.t epsilon



(a) *Mushroom* (b) *T20I10D30KP40*

Running Time w.r.t delta



(a) Varying ε (b) Varying δ

Approximation Quality in Mushroom dataset