A Probability Analysis for Candidate-Based Frequent Itemset Algorithms

Nele Dexters

University of Antwerp, Belgium

Paul P. Purdom, Dirk Van Gucht

Indiana University, Bloomington USA

Outline

- What is the paper about?
- Detailed Analysis
  - Candidate-based FIM Algorithms
  - General Shopping Model
  - Candidates
  - Probabilities
- Results
  - “Line”
  - Dataset Effects

What is the paper about?

“A probability Analysis for Candidate-Based Frequent Itemset Algorithms”

→ Theoretical Analysis of candidate generation for FIM Algorithms

→ Detailed probabilistic study of the effects of different data distributions on the performance of FIM algorithms.

Outline

- Research Area
- Analysis
  - Candidate-based FIM Algorithms
  - General Shopping Model
  - Candidates
  - Probabilities
- Results
  - Line
  - Dataset Effects

Candidate-based FIM Algorithms

- The Apriori Algorithm
- AIS
- Eclat & FP-growth
- The Fast Completion Apriori (FCA) Algorithm

General Probabilistic Shopping Model

- Identical
- Independent
- Random

$\rightarrow$ Very general: any correlation between items is possible

$\rightarrow$ Permits us to consider all sorts of data

- Uniform
- Peaky
- Anticorrelated

Candidates (1)

- An itemset is a candidate
  - No deduction of frequency status
  - Frequency has to be counted explicitly in the DB
- In practice, I is a candidate if certain associated testsets are already determined to be frequent.

Candidates (2)

Testsets

- For Apriori, AIS, Eclat & FP-growth: itemsets that are obtained by omitting a single item

- For FCA: all those subsets whose size is equal to the level where the regular Apriori Algorithm was last used.

Probabilities (1)

- Frequency status of candidate set I:
  - If I is frequent, it is called a success
  - Otherwise, it is a failure

- We study the three important corresponding probabilities:
  - Candidacy probability $C(I)$
  - Success probability $S(I)$
  - Failure probability $F(I)$

Probabilities (2)

- *C(I)= S(I) + F(I)* depends on the particular algorithm.
- All correct algorithms have the same *S(I)*
- *F(I)* depends on both the problem instance and the algorithm. It is particularly important, because it is related to work that a better algorithm might hope to avoid.

The Success Probability

$$S(I) = \sum_{j \geq k} \binom{b}{j} [P(I)]^j [1 - P(I)]^{b-j}$$

We can show that

- $P(I) \leq k/b$: $S(I) \rightarrow 0$
- $P(I) \geq k-1/b$: $S(I) \rightarrow 1$

The Failure Probability

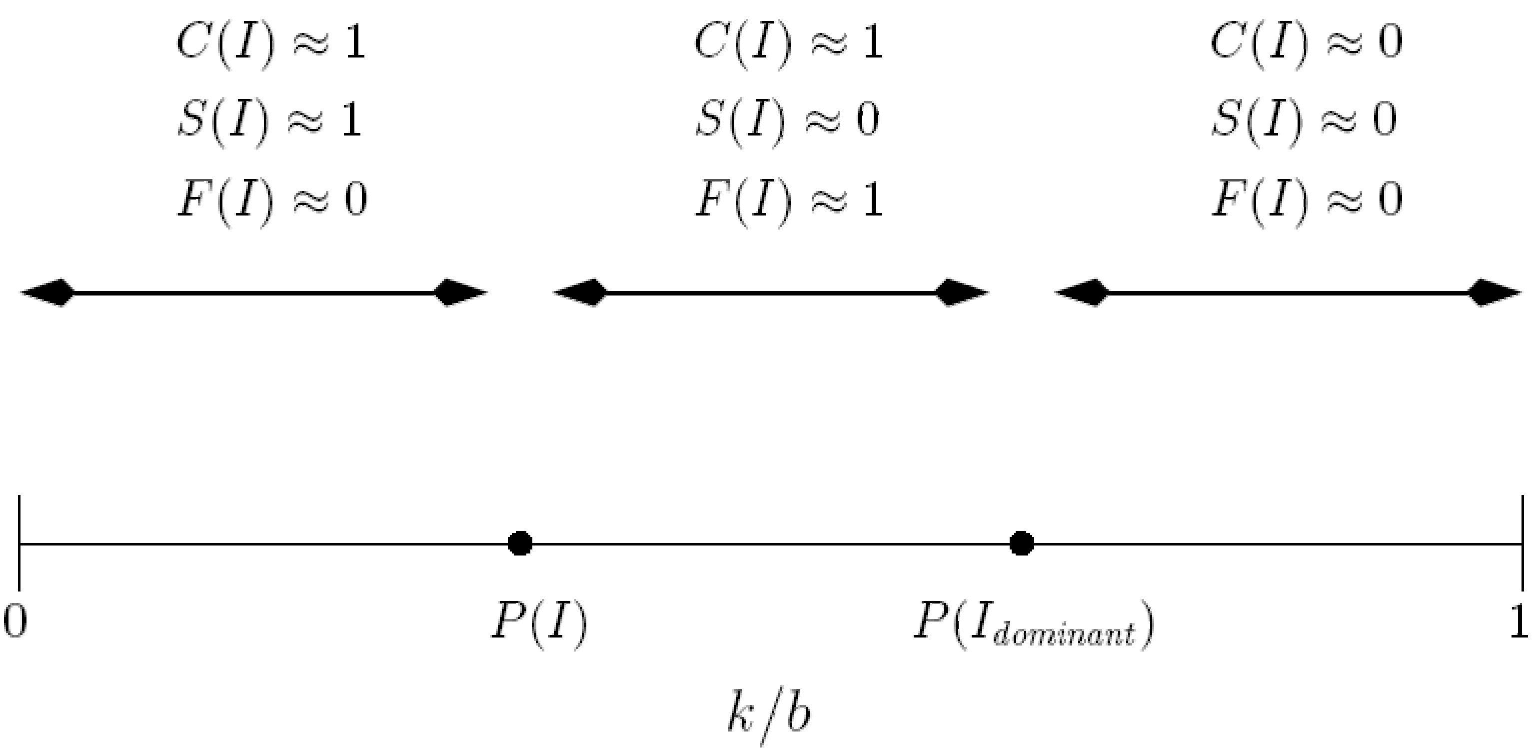
In detail for Apriori

$$
\begin{aligned}
F(I) &= C(I) - S(I) \\
&= \sum_{\substack{j_0 < k \\ j_1 \geq k - j_0 \\ j_2 \geq k - j_0 \\ \cdots \\ j_{|I|} \geq k - j_0}} \binom{b}{j_0, j_1, \ldots, j_{|I|}, b - j_0 - j_1 - \cdots - j_{|I|}} \\
&\qquad \times [P(I)]^{j_0} \left[\prod_{1 \leq i \leq |I|} Q_i(I)^{j_i} \right] \left[1 - P(I) - \sum_{1 \leq i \leq |I|} Q_i(I) \right]^{b - j_0 - \sum_{1 \leq i \leq |I|} j_i}
\end{aligned}
$$

Outline

- Research Area
- Analysis
  - Candidate-based FIM Algorithms
  - General Shopping Model
  - Candidates
  - Probabilities
- Results
  - Line
  - Dataset Effects

Line (1)


$C(I) \approx 1$
$S(I) \approx 1$
$F(I) \approx 0$
$C(I) \approx 1$
$S(I) \approx 0$
$F(I) \approx 1$
$C(I) \approx 0$
$S(I) \approx 0$
$F(I) \approx 0$
0
$P(I)$
$P(I_{dominant})$
1
k/b

Line (2)

→For both versions of the Apriori algorithm, the dominant testset is the best testset.

→For AIS, it is the worst testset.

→ For Eclat & FP-growth, the smallest of the father or the special-uncle testset controls the failure probability. This depends on the ordering that is used.

Dataset Effects

We compare the behavior of the candidate-based FIM algorithms for a variety of data distributions

General results:

- The algorithms have similar performances on uniform random data.
- The algorithms can have hugely different performance on other types of data.

Dataset Effects (2)

- Independent random data:

Apriori ≈ Eclat LFF > Eclat Lexico> Eclat MFF > AIS

- Single peak

AIS is worst

- Overlapping peaks

Apriori > Eclat LFF > Eclat Lexico> Eclat MFF > AIS

- For anticorrelated data, it is even possible to have Eclat MFF > Eclat LFF

Thank you!