Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees

Matteo Riondato and Eli Upfal
Department of Computer Science – Brown University
matteo@cs.brown.edu

I. Settings and definitions

**Transactional Dataset D**

<table>
<thead>
<tr>
<th>Itemset</th>
<th>Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bread, Milk</td>
<td>1</td>
</tr>
<tr>
<td>Bread, Diaper, Beer, Eggs</td>
<td>2</td>
</tr>
<tr>
<td>Milk, Diaper, Beer, Coke</td>
<td>3</td>
</tr>
<tr>
<td>Bread, Milk, Diaper, Beer, Coke</td>
<td>5</td>
</tr>
</tbody>
</table>

**Formula**: Frequency of an itemset X in D: \( f_{D}(X) = \text{fraction of transactions of D containing X} \)

**Association Rule**: \( X \rightarrow Y \)

- "transactions containing X are likely to contain Y" (strength)
- frequency of \( X \rightarrow Y \) is \( f_{D}(X \cup Y) / f_{D}(X) \) (confidence)

**Mining Problems**: Find the sets

1. Frequent Itemsets with threshold \( \theta \)
   - All itemsets X with frequency \( f_{D}(X) \geq \theta \)
2. Top-K Frequent Itemsets
3. All itemsets at least as frequent as the Kth frequent

**Example**: This dataset has d-index \( d=3 \)

**Notations**

- **D**: Transactional Dataset
- **S**: Sample
- **R**: Frequent Itemsets
- **\( \theta \)**: Confidence
- **\( \delta \)**: Frequency accuracy

**Algorithm**

1. Compute |S| and create S using random sampling with replacement
2. Output |S| and create S using random sampling with replacement

**Theorem**: Correctness

The set \( \mathbb{F}(S, \theta - \varepsilon/2) \) is an \( \varepsilon \)-approximation to \( \mathbb{F}(D, \theta) \) with probability at least \( 1 - \delta \)

**V. The d-index of the dataset**

**Definition**: The d-index \( d \) of a dataset D is the maximum integer such that D contains at least \( d \) transactions of length at least \( d \).

**Example**: This dataset has d-index \( d=3 \)

The d-index can be computed with a single scan of the dataset.

**Theorem**: \( d \) is an upper bound to the VC-Dimension of \( \mathbb{F}(D, \theta) \).

**II. Motivation, Goals and Constraints**

**Exact algorithms exist** for the mining problems (Apriori, FP-Growth,...)

They have drawbacks:
- Need to scan dataset D multiple times
- Running time depends on size of D (number of transactions)
- Too expensive for very large datasets: disk access is slow

**Goal**: Speed up mining using a random sample of D while guaranteeing good results

**Sample**: collection of transactions drawn uniformly at random from D

**Constraints**

- Sample should fit in main memory: no disk access
- Size of sample must not depend on size of D
- Given probabilistic guarantees on quality of the results.
- Make no assumptions on the frequencies distribution.

**IV. VC-Dimension**

**Tool from Statistical Learning Theory**

- Describes "richness" of family of indicator functions
- Gives bound to sample size needed to approximately learn a function

**Definition**

Given a set of points \( P \) and a family \( F \subseteq 2^P \) (ranges), the **VC-Dimension** of the range space \( (P, F) \) is the cardinality of the largest \( A \subseteq P \) such that \( \left| r \cap A : r \in F \right| = 2^d \)

**Theorem**: Bound to sample size

Given \( \delta < 1 \), if \( (P, F) \) has VC-Dimension \( 5d \), with probability \( 1 - \delta \), a random sample \( S \subseteq P \) of size \( |S| \geq \left\lceil \frac{4}{\delta} \right\rceil (4d + \log \frac{1}{\delta}) \)

is such that

- \( |\mathbb{F}(S, \theta - \varepsilon/2)| \leq \varepsilon/|F| \) \( \forall F \in F \)

**Results**

- Sample always fits in main memory (hundreds of runs)
- \( \mathbb{F}(S, \theta - \varepsilon/2) \) always an \( \varepsilon \)-approximation to \( \mathbb{F}(D, \theta) \)
- Frequency accuracy even better than guaranteed
- Mining time significantly improved

**VI. Experiments**

We evaluated our method using datasets from FIMI repository

**Results**

- Sample always fits in main memory (hundreds of runs)
- \( \mathbb{F}(S, \theta - \varepsilon/2) \) always an \( \varepsilon \)-approximation to \( \mathbb{F}(D, \theta) \)
- Frequency accuracy even better than guaranteed
- Mining time significantly improved