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

TID	Items
1	Bread, Milk
2	Bread, Diaper, Beer, Eggs
3	Milk, Diaper, Beer, Coke
4	Bread, Milk, Diaper, Beer
5	Bread, Milk, Diaper, Coke

Transaction
Items
Itemset

Figure from Tan et al. - *Introduction to Data Mining*

Frequency of an itemset X in D:
$f_D(X)$= fraction of transactions of D containing X

Association Rule W: X→Y
- "transactions containing X are likely to contain Y"
- frequency of W: $f_D(W) = f_D(X \cup Y)$ **confidence** of W: $c_D(W) = \frac{f_D(X \cup Y)}{f_D(X)}$

Mining Problems: Find the sets
- $\mathrm{FI}(D, \theta)$: **Frequent Itemsets** with threshold
  - All itemsets X with frequency $f_D(X) \geq \theta$, with their frequencies in D
- $\mathrm{TOPK}(D, K)$: **Top-K Frequent Itemsets**
  - All itemsets at least as frequent as the Kth most frequent
- $\mathrm{AR}(D, \theta, \gamma)$: **Association Rules**
  - All association rules W with $f_D(W) \geq \theta$ and $c_D(W) \geq \gamma$

Our work can be applied to all three problems

II. Motivation, Goals and Constraints

Exact algorithms exist for the mining problems (Apriori, FPGrowth,...)

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

Key Observation: Data mining is **exploratory** in nature
Fast and good enough results are preferred to slow but exact

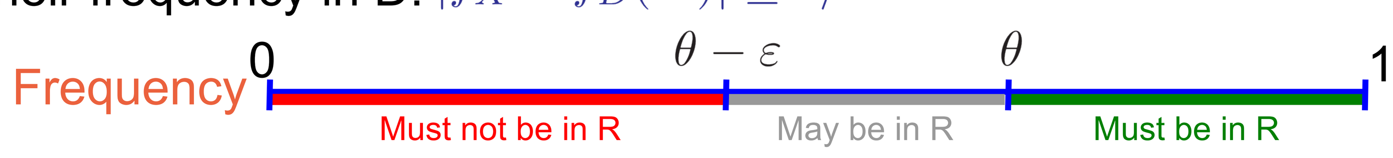
Goal: Speed up mining using a **random sample** of D while **guaranteeing good results**
Sample = **collection of transactions** drawn **uniformly at random** from D

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

III. Our solution

(focus on $\mathrm{FI}(D, \theta)$. Everything can be extended to the other problems)

Desired **properties of the output:** a set R of pairs (X, f_X) such that
- All itemsets with frequency $f_D(X) \geq \theta$ must be in R.
- No itemset with frequency $f_D(X) < \theta - \varepsilon$ can be in R.
- All itemsets X in R must have an associated f_X close (within ε/2) to their frequency in D: $|f_X - f_D(X)| \leq \varepsilon/2$



R=ε**-approximation to** $\mathrm{FI}(D, \theta)$
Variant with **relative guarantees** ($(1-\varepsilon)\theta$, ...) in the paper

Key Ingredient: Use results on **VC-Dimension** to compute |S| such that for a sample S of size |S|, we have:

$$\boxed{\Pr(\exists \text{ itemset } X \ : \ |f_D(X) - f_S(X)| > \frac{\varepsilon}{2}) < \delta}$$

Algorithm
input: D, $\theta, \varepsilon, \delta$
1) Compute |S| and create S using random sampling with replacement
2) Output $\mathrm{FI}(S, \theta - \varepsilon/2)$ using exact algorithm

Theorem: Correctness
The set $\mathrm{FI}(S, \theta - \varepsilon/2)$ is an ε-approximation to $\mathrm{FI}(D, \theta)$ with probability at least $1 - \delta$

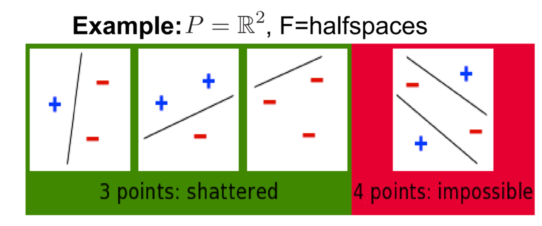
IV. VC-Dimension

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

$$\boxed{\{r \cap A \ : \ r \in F\} = 2^A}$$

Example: $P = \mathbb{R}^2$, F=halfspaces



Theorem: Bound to sample size
Given 0<ε, δ< 1, if (P,F) has **VC-Dimension ≤d**, with **probability** $\geq 1 - \delta$, a random sample $S \subseteq P$ of size

$$\boxed{|S| \geq \frac{1}{\varepsilon^2}\left(d + \log\frac{1}{\delta}\right)}$$

If d does not depend on |D|, then **|S| is also independent from |D|!**

is such that

$$\boxed{\left|\frac{|f|}{|P|} - \frac{|S \cap f|}{|S|}\right| \leq \varepsilon, \forall f \in F}$$

Simultaneous deviation bound on all the ranges!

Such a sample is called an ε**-approximation to (P,F)**

V. The d-index of the dataset

In our case:
- $P = D$
- For any itemset X, let $T_D(X)$ = set of transactions of D containing X
- $F_D = \{T_D(X) \forall \text{ itemset } X\}$

If $S \subseteq D$ is an ε**/2-approximation** for (D, F_D):

$$\left|\frac{|T_D(X)|}{|D|} - \frac{|S \cap T_D(X)|}{|S|}\right| \leq \frac{\varepsilon}{2}, \forall \text{ itemset } X$$

i.e. $\boxed{|f_D(X) - f_S(X)| \leq \varepsilon/2, \forall \text{ itemset } X}$

We need a bound to the VC-Dimension of (D, F_D)

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** ← **d is independent from |D|**

Example: this dataset has d-index d=3

bread	beer	milk	coffee
chips	coke	pasta	
bread	coke	chips	
milk	coffee		
pasta	milk		

The d-index can be computed with a single scan of the dataset
Theorem: d is an **upper bound to the VC-Dimension** of (D, F_D)
Theorem: there are datasets with VC-Dimension exactly d
i.e., the **bound is strict**

VI. Experiments

We evaluated our method using datasets from FIMI repository

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