Sampling for Frequent Itemset Mining
The Power of PAC learning

prof. dr Arno Siebes

Algorithmic Data Analysis Group
Department of Information and Computing Sciences
Universiteit Utrecht

What We'll Do Today

We discuss

- the regular content: Riondato and Upfal on sampling
  - the reason why we discussed PAC learning
- as well as the essay

Helpful material for your essay is also – and importantly – on the website: cs.uu.nl/docs/vakken/mdb

Please, read this and use it.

The Papers

Today we discuss:

Efficient Discovery of Association Rules and Frequent Itemsets through Sampling with Tight Performance Guarantees

- Proc ECML PKDD 2012, LNCS 7523
- ACM Transactions on Knowledge Discovery from Data, Vol 8, No 4, 2014

Note, we restrict ourselves to frequent itemset parts only

- the association rules parts are also interesting, but beyond our scope

The Problem

Let D be a transaction database over $\mathcal{I}$ (which we assume fixed throughout the lecture). Denote by $F(D, \theta)$ the set of frequent itemsets and their support, i.e.,

- $(I, supp_D(I)) \in F(D, \theta) \Leftrightarrow supp_D(I) \geq \theta$.

We want to compute a (small) sample $S \subset D$ such that

- $F(S, \theta') \approx F(D, \theta)$

More formally, we want that our sample to yield an (ϵ, δ) approximation:

For $(\epsilon, \delta) \in (0, 1)^2$, an (ϵ, δ) approximation of $F(D, \theta)$ is a set $C = \{(I, s(I)) \mid I \subseteq \mathcal{I}, s(I) \in (0, 1]\}$ such that with probability at least $1 - \delta$:

1. $(I, supp_D(I)) \in F(D, \theta) \rightarrow (I, s(I)) \in C$
2. $(I, s(I)) \in C \rightarrow supp_D(I) \geq \theta - \epsilon$
3. $(I, s(I)) \in C \rightarrow |supp_D(I) - s(I)| \leq \epsilon/2$

Now in Natural Language

C is an (ϵ, δ) approximation of $F(D, \theta)$ if with probability at least $1 - \delta$:

- C contains all frequent itemsets
- the non-frequent itemsets in C are almost frequent
- our estimate of the frequency of the itemsets in C is approximately correct.

This means that (with high probability)

- C may contain false positives
- but *no* false negatives
- and there is limited loss of accuracy

Which means that we can compute $F(D, \theta)$

- with *one* scan over D using C.

In Terms of Samples

In the terminology of samples we can rephrase our goal now as

Find a sample S such that

$$\mathbb{P}\left(\exists I \subseteq \mathcal{I} : |supp_D(I) - supp_S(I)| > \epsilon/2\right) < \delta$$

Because if this holds for S, then

- $F(S, \theta - \epsilon/2)$ is an (ϵ, δ) approximations of $F(D, \theta)$

For the simple reasons that

1. with probability at least $1 - \delta$: $F(D, \theta) \subset F(S, \theta - \epsilon/2)$
2. $I \in F(S, \theta - \epsilon/2) \rightarrow$
   [with prob at least $1 - \delta$: $supp_D(I) \geq supp_S(I) - \epsilon/2 \geq \theta - \epsilon$]
3. $I \in F(S, \theta - \epsilon/2) \rightarrow$
   [with prob at least $1 - \delta$: $|supp_D(I) - supp_S(I)| \leq \epsilon/2$]

Classifiers and Hypothesis Sets

A classifier is a simply a function

$$f : X \to \{0, 1\}$$

A set of hypotheses $\mathcal{H}$

- is simply a set of classifiers, i.e., set set of such functions

There is a 1-1 relation between classifiers and subsets of X:

- each subset A of X has an indicator function $\mathbb{1}_A$:

$$\mathbb{1}_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}$$

- which is a classifier on X

In other words, we can see $\mathcal{H}$

- as a set of subsets of X

Range Space

This is formalised as a *range space*

A range space (X, R) consists of

- a finite or infinite set of points X
- a finite of infinite family R of subsets of X, called ranges

For any $A \subset X$, the projection of R on A is given by

$$\Pi_R(A) = \{r \cap A \mid r \in R\}$$

So, in this terminology, we have that

- a subset $A \subset X$ is shattered by R if $\Pi_R(A) = P(A)$

ϵ-Approximations

ϵ representative samples translate to ϵ- approximations:

Let (X, R) be a range space and let $A \subset X$ be a finite subset. For $\epsilon \in (0, 1)$, a $B \subset A$ is an ϵ-approximation for A if

$$\forall r \in R : \left| \frac{|A \cap r|}{|A|} - \frac{|B \cap r|}{|B|} \right| \leq \epsilon$$

without proof (it should look and feel familiar):

There is a constant c (≤ 0.5 experimentally) such that if (X, R) is a range space of VC dimension at most υ, $A \subset X$ a finite subset, $(\epsilon, \delta) \in (0, 1)^2$, then a random $B \subset A$ of cardinality

$$m \geq \min \left\{ |A|, \frac{c}{\epsilon^2} \left(\upsilon + \log \frac{1}{\delta} \right) \right\}$$

is an ϵ-approximation for A with probability at least $1 - \delta$

A Range Space for Transaction Data

To use this bound, we should first formulate our hypothesis set, i.e., our range space:

Let D be a transaction database over $\mathcal{I}$. The associated range space $S = (X, R)$ is defined by

1. $X = D$, the set of transactions
2. $R = \{T_D(I) \mid I \subseteq \mathcal{I}, I \neq \emptyset\}$, where $T_D(I) = \{t \in D \mid I \subseteq t\}$
   - the range of I, $T_D(I)$, is the set of transactions that support I
   - i.e., its support set.

Note that R contains exactly the support sets of the closed itemsets

- the support set of a non-closed itemset is the support set of a closed itemset as well.

Now all we have to do is to determine the VC dimension of this range space.

- as more often, we'll settle for a (tight) upper bound.

A First Result

Let D be a dataset with associated range space $S = (X, R)$. Then $VC(S) \geq \upsilon \in \mathbb{N}$ if there exists an $A \subset X$ with $|A| = \upsilon$ such that

$$\forall B \subset A : \exists I_B \subset \mathcal{I} : T_A(I_B) = B$$

$\Leftarrow$: For all $B \subset A$, $T_D(I_B) \cap A = B$, that means that $B \in \Pi_R(A)$ for all B. That means that $\Pi_R(A) = P(A)$. A is shattered and, thus, $VC(S) \geq \upsilon$.

$\Rightarrow$: If $VC(S) \geq \upsilon$ then there is an $A \subseteq D$ with $|A| = \upsilon$ such that $\Pi_R(A) = P(A)$. Hence, for every $B \subseteq A$, there is an itemset I_B such that $T_D(I_B) \cap A = B$. Since $T_A(I_B) \subseteq T_D(I_B)$, this means that $T_A(I_B) \cap A \subseteq B$. Now note that

- $T_A(I_B) \subseteq A$ and
- by construction $B \subseteq T_A(I_B)$

and thus $T_A(I_B) = B$

An Immediate Consequence

Let D be a dataset with associated range space $S = (X, R)$. Then $VC(S)$ is the largest $\upsilon \in \mathbb{N}$ such that there is an $A \subseteq D$ with $|A| = \upsilon$ such that

$$\forall B \subset A : \exists I_B \subset \mathcal{I} : T_A(I_B) = B$$

Example

$D = \{\{a, b, c, d\}, \{a, b\}, \{a, c\}, \{d\}\}$ and $A = \{\{a, b\}, \{a, c\}\}$

- $I_{\{\{a,b\}\}} = \{\{a, b\}\}$
- $I_{\{\{a,c\}\}} = \{\{a, c\}\}$
- $I_{\{\{a\}\}} = A$
- $I_{\emptyset} = \{\{d\}\}$

Larger subsets of D cannot be shattered and hence its VC dimension is 2.

Nice, But

It is good to have a simple characterisation of the VC dimension. But since it puts a requirement on:

- $\forall B \subset A$

it is potentially very costly to compute

- in fact, it is known to be $O(|R||X|^{\log |R|})$

Fortunately, our corollary (the immediate consequence) suggests an alternative

- we need a set of υ transactions
- if all of them are at least υ long

we have enough freedom to make the condition hold

- for technical reasons we first assume that the transactions are independent, i.e., they form an antichain.

The d-index

Let D be a data set. The *d-index* of D is the largest integer d such that

- D contains at least d transactions of length at least d
- that form an antichain

Theorem
Let D be a dataset with d-index d. Then the range space $S = (X, R)$ associated with D has VC dimension of at most d

This upper bound is tight

- there are datasets for which the VC dimension equals their d-index

Proof Sketch

Let $l > d$ and assume that $T \subset D$ with $|T| = l$ can be shattered

- note that this means that T is an antichain: if $t_1 \subseteq t_2$ all ranges containing t_2 also contain t_1: we cannot shatter

For any $t \in T$, there are 2^{l-1} subsets of T that contain t. So, t occurs in 2^{l-1} ranges T_A.
t only occurs in T_A if $A \subset t$. Which means that T occurs in $2^{|t|} - 1$ ranges.
From the definition of d we know that T must contain a t^* such that $|t^*| < l$

- otherwise the d-index would be l

This means that $2^{|t^*|} - 1 < 2^{l-1}$, so t^* cannot appear in 2^{l-1} ranges.
This is a contradiction. So, the assumed T cannot exist. Hence, the largest set that is shattered has at most size d.

From d-Index to d-Bound

The d-index of D is still a bit hard to compute

- because of the antichain requirement

So, lets us forget about that requirement.

Let D be a dataset, its *d-bound* of D is the largest integer d such that

- D contains at least d different transactions of length at least d

Theorem
Let D be a dataset with d-bound d. Then the range space $S = (X, R)$ associated with D has VC dimension of at most d

This is obvious as *d-bound* $\geq$ *d-index*

- a subset witnessing the d-index satisfies the conditions for the d-bound (but not vice versa)

Computing The d-Bound

Computing the d-bound is easy

- do a scan of the dataset
- maintaining
  - the ℓ longest (different) transactions
  - that are at least ℓ long
  - breaking ties arbitrarily

See the journal version for the full details

- and the proof!

The Sample Size (finally)

Combining all the results we have seen so far, we have:

Let D be a dataset, let d be the d-bound of D, and let $(\epsilon, \delta) \in (0,1)^2$. Let S be a random sample of D with size

$$|S| \geq \min \left\{ |D|, \frac{4c}{\epsilon^2} \left(d + \log \frac{1}{\delta} \right) \right\}$$

Then $F(S, \theta - \epsilon/2)$ is an ϵ approximation of $F(D, \theta)$ with probability at least $1 - \delta$.

Such a sample we can easily compute from D in a single scan. Hence we need two scans of D to compute a ϵ approximation of the set of all frequent itemsets.

Your Essay

The essay you have to submit consists of

1. an explanation of the results we achieved today: 5 - 6 pages

You have to submit by April 11, 9AM

You submit by email, to me.

- subject of email contains: [ESSAY BIG DATA], your name and your student number
  - automatic processing then ensures that I will see and grade your essay
- provide the same information at the start of your essay
  - to ensure that I know who I should assign the grade to.
- Using your name and student number in the name of the file you submit is a nice gesture.

Retake: submit by July 8, 12 midnight

Writing an Essay

Most of you did not submit essays before. The most important guideline is:

- Be coherent: define the concepts (terms) you use and use them coherently
- define before you use

On writing:

- Use a spell checker and check whether or not a word or an expression you use means what you think it means.
- You have to explain non-trivial material. Long sentences are more confusing, so keep your sentences short
- Paragraphs have 1 message only, multiple messages means multiple paragraphs.
- Sections are a top-level division of your argument, use this to guide your reader.

Page Limits

The page limits are strict

- violations will cost you dearly

The reason is twofold

- If you cannot explain in the allotted number of pages, you probably do not understand
- unlimited number of pages would make marking impossible

You are free

- to choose your favourite text processor: troff, word, latex, ...
- given that you need graphs, tables, math, latex might be a wise choice
- we provide you a .cls file
- if you use another text processor, please emulate this style

Content

Very briefly one could say that all you have to do is:

- recursively explain today's lecture

and a bit more ...

There are detailed instructions,

- available as a pdf file
- but also as a latex file, which you can use as a template

Since your grade depends on how well your essay answers the questions raised by this document

- it seems wise to use the template as an outline of your essay!