Mining Meaningful Patterns

Konstantin Tretyakov, 4.11.2009
So far

- Frequent, maximal, closed itemsets.
- Fast enumeration algorithms
- Missing values & noise
- Swap randomization
Theorem 2. Consider an asymptotic regime where as $n \to \infty$, we have $k, s = O(1)$ with $s \geq 2$, $E[R^{2s}] = O(n^{-a})$ for some constant $2 < a \leq 2s$, and $t = O(n^c)$ for some positive constant $c$. If
\[
ce \leq \frac{(k - 1)(a - 2) + \min(2a, 6, 0)}{2s},\]
then the variation distance between the distributions $\mathcal{L}(\hat{Q}_{k,s})$ and $\mathcal{L}(U)$ of $\hat{Q}_{k,s}$ and $U$ satisfies
\[
\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\| = \sup_A |\Pr(\hat{Q}_{k,s} \in A) - \Pr(U \in A)| = O(1/n).
\]

Proof. Applying Theorem 1 gives
\[
\|\mathcal{L}(\hat{Q}_{k,s}) - \mathcal{L}(U)\| \leq b_1 + b_2
\]
where
\[
b_1 = \sum_{X:|X|=k} \sum_{Y \in I(X)} p_{XY}E[Z_X Z_Y].
\]
and
\[
b_2 = \sum_{X:|X|=k} \sum_{Y \neq X \in I(X)} E[Z_X Z_Y].
\]

We now evaluate $b_1$ and $b_2$. Letting $\hat{R}$ denote the vector of the $R_x$’s, we have that for any set $X$ of $k$ items
\[
\mathbb{E}[Z_X Z_Y | \hat{R}] \leq \sum_{i=0}^{s} \binom{t}{i, s-i, s-i} \left( \prod_{x \in X \cup Y} R_x^{i} \right) \times \left( \prod_{x \in X} R_x^{s-i} \right) \left( \prod_{y \in Y} R_y^{s-i} \right).
\]

Applying independence of the $R_x$’s and Jensen’s inequality gives
\[
\mathbb{E}[Z_X Z_Y] = \mathbb{E}[\mathbb{E}[Z_X Z_Y | \hat{R}]]
\]
\[
\leq \sum_{i=0}^{s} \binom{t}{i, s-i, s-i} \mathbb{E}[R^{2s-i}] \mathbb{E}[R^{i}]^{2(k-g)}
\]
\[
\leq \sum_{i=0}^{s} t^{2s-i} \mathbb{E}[R^{2s}]^{\frac{g(2s-i)}{2s}} \mathbb{E}[R^{2s}]^{k-g}
\]
\[
= \sum_{i=0}^{s} t^{2s-i} \mathbb{E}[R^{2s}]^{k-i/2s}.
\]

Today

- Significant itemsets
- Non-redundant itemsets
- Absolutely significant itemsets
Significantly frequent itemsets
Expected counts

Assume all items are independent

\[ P(\text{red}) = \frac{4}{5} = 0.8 \]

\[ P(\text{green}) = \frac{4}{5} = 0.8 \]

\[ P(\text{red and green}) = 0.8 \times 0.8 = 0.64 \]

\[ P(3 \times \text{red and green out of 5}) = 10 \times 0.64^3 \times (1 - 0.64)^2 = 0.34 \]
In general

\[ S = \{1, 3, 6, 8\} \]

\[ p_S = p_1 p_3 p_6 p_8 \]
In general

- $S = \{1, 3, 6, 8\}$
- $P_S = P_1 P_3 P_6 P_8$
- $P(S \text{ is present } k \text{ times out of } n) = \binom{n}{k} P_S^k (1 - P_S)^{n-k}$
In general

- \( S = \{1,3,6,8\} \)
- \( P_S = P_1 \cdot P_3 \cdot P_6 \cdot P_8 \)
- \( P(S \text{ is present } k \text{ times out of } n) = \)
  \[
  = \binom{n}{k} P_S^k (1 - P_S)^{n-k}
  \]
- \( P(S \text{ is present } k \text{ or more times}) = \)
  \[
  = \sum \binom{n}{k} P_S^k (1 - P_S)^{n-k}
  \]

\( P\)-value\( (S) \)
Redundancy

*Redundant* significantly frequent itemsets

1
2
3
4
5
Redundancy

- $P$-value($S \text{ given } U$)

- Probability to see itemset $S$ in random data at least $k$ times.
Redundancy

- **P-value**(*S* given *U*):
  - Probability to see itemset *S* in random data at least *k* times.

Assume all items are **independent**.
Redundancy

- P-value($S \g U$)
  - Probability to see itemset $S$ in random data at least $k$ times.

Assume $m$ transactions contain itemset $U$, and otherwise items are independent.

Assume all items are independent.
Redundancy

\[ p_{I}^{(1)} = P(I \subseteq is(t)) = P(i \in I | c1)P(c1) + P(i \in I | c2)P(c2), \]

\[ = \left( 1 - \frac{supp(I_1)}{n} \right) \prod_{i \in I} p_i + \frac{supp(I_1)}{n} \prod_{i \in I \setminus I_1} p_i. \]

\[ p_{I}^{(1)} = \frac{1}{n} \sum_{t \in \mathcal{T}} \left( \prod_{i \in I \setminus \mathcal{I}^{(1)}(t)} p_i \right) \]
Compute P-values for all itemsets.

Output the itemset with the lowest P-value.

Recompute other P-values to accommodate the newly covered itemset.

* this can be highly optimized.

Repeat

(Gallo, Mammone, De Bie, Turchi & Cristianini, 2007)
But what P-values are “good”?

- Let $S = \text{color circles}$

- Let $\text{supp}(S) = 20$ and $P\text{-value}(S) < 0.05$.
  - i.e. probability of having $S$ at least 20 times in random data is $< 5\%$.

- Does it mean that $S$ is significant?
Hypothesis testing perspective

Consider an algorithm:

\[ P\text{-value}(S) < 0.05 \implies S \text{ is significant} \]

On random data, you’ll get a false positive with probability 5%
Multiple testing

- $P$-value( ) = 0.11
- $P$-value( ) = 0.32
- $P$-value( ) = 0.04
- $P$-value( ) = 0.59
- $P$-value( ) = 0.21
- $P$-value( ) = 0.15
- $P$-value( ) = 0.26
- $P$-value( ) = 0.01

...
Multiple testing

- $P$-value( ) = 0.11
- $P$-value( ) = 0.32
- $P$-value( ) = 0.04
- $P$-value( ) = 0.59
- $P$-value( ) = 0.21
- $P$-value( ) = 0.15
- $P$-value( ) = 0.26
- $P$-value( ) = 0.01

...
Bonferroni correction

- Suppose we make a total of \( N \) tests.
- Consider an algorithm:

\[
P\text{-value}(S) < \frac{0.05}{N} \Rightarrow S \text{ is significant}
\]

- On random data, each of this tests is a false positive with probability < 0.05/N
- Therefore the probability to have at least one false positive among all \( N \) is < 0.05
Bonferroni correction

- So what is N for frequent itemset mining?
- How much should you adjust the P-value?
- Can we do better?
False Discovery Rate

- Suppose you make 100 tests.

- Suppose that all of them have $P$-value $< 0.05$.

- However, suppose that none of them has $P$-value $< 0.0005$. 
False Discovery Rate

- Do not ban all false positives.
- Allow some, but at most 5% of all discoveries.
False Discovery Rate

- Do not ban all false positives.
- Allow some, but at most 5% of all discoveries.
- Not enough to decide at the level of individual P-values.
  - Instead, compare the number of discoveries with the expected number for various thresholds and choose.

E.g. Simes procedure (Benjamini & Hochberg):
- Find largest k such that $P_{(k)} < 0.05 \frac{k}{N}$
- Declare all tests 1..k as significant.
Efficient FDR for frequent itemsets

An Efficient Rigorous Approach for Identifying Statistically Significant Frequent Itemsets

Adam Kirsch
Harvard University

Michael Mitzenmacher
Harvard University

Andrea Pietracaprina
University of Padova

Geppino Pucci
University of Padova

Eli Upfal
Brown University

Fabio Vandin
University of Padova
Summary

- Random data, statistical significance, P-values
- Accounting for redundancy
  - MINI
- Finding “truly significant” itemsets.
  - Bonferroni, FDR, ++
Questions?

Why are we doing this?
What problem are we solving?
Is this actually useful?
Are we adding value?
Will this change behavior?
Is there an easier way?
What's the opportunity cost?
Is it really worth it?