Probably the best itemsets

Bayesian approach for ranking itemsets

Nikolaj Tatti

nikolaj.tatti@gmail.com

ADReM, Universiteit Antwerpen, Antwerpen, Belgium

Problem of Pattern Explosion

- Pattern explosion is the biggest setback in pattern mining.
- Rank/prune the itemsets by comparing the observed support against the expected value.
  - Large difference in supports = interesting pattern.
- An independence model is a popular choice.

Why this is bad?

- We discover the same information multiple times.
- Consider a data set with K items such that
  - $a_1 = a_2$
  - the rest of items are independent.
- Any itemset containing both a_1 and a_2 does not follow independence assumption.
- There will be 2^{K-2} interesting itemsets.
- However, to explain the data we need to know only the frequencies of singletons and $a_1 a_2$.

Pattern set mining

- Recent trend in pattern mining.
- Score a pattern set as a whole instead of single pattern.
- By doing so can remove redundancy more efficiently.
- Statistical approach:
  - Build a statistical model from the current patterns.
  - Fit the model into data,

    model explains data well = current pattern set is good.
  - Pattern set selection = model selection.
- Use heuristics to find a good collection.

Can we use pattern set measures for scoring individual itemsets?

Recipe for Scoring Itemsets

- You need
  - a set of statistical models, say $M_1, \ldots, M_K$,
  - a function *fam* mapping a model M_i to some *downward closed* itemset collection, $F_i = fam(M_i)$.
- $p(M_i \mid D)$ is the posterior probability of the ith model.
- Score of an itemset X

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

Example

- Assume 3 models,

Model	Itemsets	$p(M \mid D)$
M_1	a, b, c, d, ab, bc, cd	0.5
M_2	a, b, c, d, ab, ad	0.3
M_3	a, b, c, d, bc, cd	0.2

Example

- Assume 3 models,

Model	Itemsets	$p(M \mid D)$
M_1	a, b, c, d, ab, bc, cd	0.5
M_2	a, b, c, d, ab, ad	0.3
M_3	a, b, c, d, bc, cd	0.2

- The scores for singletons are

$$sc(a) = sc(b) = sc(c) = sc(d) = 1,$$

- The scores for non-singletons are

$$sc(ab) = 0.8, sc(bc) = 0.7, sc(ad) = 0.3, sc(cd) = 0.7.$$

Scoring Itemsets

- The score decreases monotonically but...

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

Scoring Itemsets

- ...we have 4 problems.

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

Scoring Itemsets

- ...we have 4 problems.

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

- How to define M_i?

Scoring Itemsets

- ...we have 4 problems.

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

- How to define M_i?
- How to define $F_i = fam(M_i)$?

Scoring Itemsets

- ...we have 4 problems.

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

- How to define M_i?
- How to define $F_i = fam(M_i)$?
- How to compute the probability $p(M_i \mid D)$?

Scoring Itemsets

- ...we have 4 problems.

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D).$$

- How to define M_i?
- How to define $F_i = fam(M_i)$?
- How to compute the probability $p(M_i \mid D)$?
- How to compute the sum?

Defining Model

- Use exponential models (a.k.a log-linear or maximum entropy models):
  - The mapping *fam* will be natural.
  - Connections with maximum entropy principle.
  - Connections with MDL theory.
  - Empirical demonstrations for being a good estimate.
- Posterior $p(M \mid D)$ can be estimated for a large subset of exponential models.

Why exponential model is so great

- If M is the simplest model (smallest $|fam(M)|$) that explains the data, then
  - $sc(X) \to 1$ if $X \in fam(M)$.
  - $sc(X) \to 0$ if $X \notin fam(M)$.

Computing the sum

- Instead of computing

$$sc(X) = \sum_{X \in F_i} p(M_i \mid D)$$

sample N models from $p(M \mid D)$ and estimate

$$sc(X) \approx \frac{\text{number of models for which } X \in \mathit{fam}(M)}{N}.$$

- Use MCMC.

Some examples

- Course enrollment data for CS courses in Helsinki.
- 4 most interesting (non-singletons) itemsets
  - *Computer Architectures*, *Performance Analysis* (0.95)
  - *Design and Analysis of Algorithms*, *Principles of Functional Programming* (0.94)
  - *Database Systems II*, *Information Storage* (0.94)
  - *Three concepts: probability*, *Machine Learning* (0.92)

That's it!