Probably the Best Itemsets

nikolaj.tatti@ua.ac.be

Universiteit Antwerpen

Pattern Explosion

Pattern explosion is the biggest setback in pattern mining. A common approach to solve this is to rank/prune the itemsets by comparing the observed support against the expected value, say, w.r.t. independence assumption,

difference in supports = interesting pattern.

The problem is that we discover the same information multiple times. For example, consider a data set with K items:

- $a_1 = a_2$
- the rest of items are independent.

Any itemset containing both a_1 and a_2 does not follow independence assumption $\rightarrow$ there will be 2^{K-2} interesting itemsets.
However, to explain the data we need to know only the frequencies of singletons and a_1a_2.

Pattern Set Mining

To reduce the redundancy, score *itemset collections* instead of ranking single itemsets.
Statistical approaches:

- Let $\mathcal{F}$ be an itemset collection.
- Build a statistical model M from a $\mathcal{F}$.
- Fit the model into data

$$M \text{ explains data well} = \mathcal{F} \text{ is good.}$$

- Pattern set selection = model selection.

Heuristics are used to find a good pattern set.

Score

Use measures for pattern sets to score individual itemsets.

You need

- a set of models, say $M_1, \ldots, M_K$,
- a function *fam* mapping a model M_i to some *downward closed* itemset collection, $\mathcal{F}_i = \mathit{fam}(M_i)$.

Score of an itemset X

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

where $p(M_i \mid D)$ is posterior probability of the ith model.

Toy 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 are

$$sc(a) = sc(b) = sc(c) = sc(d) = 1,$$
$$sc(ab) = 0.8, sc(bc) = 0.7,$$
$$sc(ad) = 0.3, sc(cd) = 0.7.$$

Exponential Models

Exponential models provide natural set of models.

- The mapping fam will be natural.
- Connections with maximum entropy.
- Connections with MDL theory.
- Empirical demonstrations for being a good estimate.

Let $\mathcal{F}$ be a (downward closed) collection of itemsets. Exponential model M is defined as

$$p(t \mid r, M) = \exp\left(\sum_{X \in \mathcal{F}} r_X S_X(t)\right),$$

where r_X is a parameter and $S_X(t) = 1$ iff X covers t. Define $F = fam(M)$.

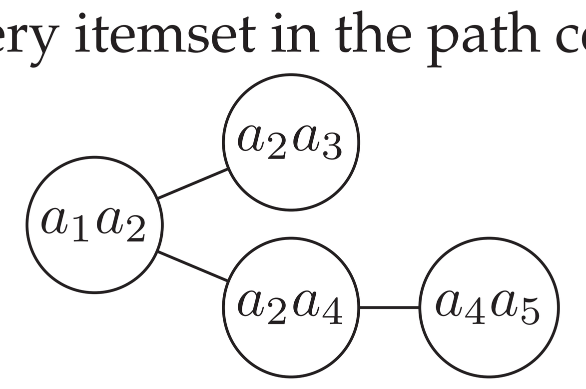
Decomposable Models

The posterior is proportional to

$$p(M \mid D) = \text{bayes' tricks} \propto \prod_{t \in D} \int_r p(t \mid r, M).$$

Estimate integral with a BIC score. BIC score cannot be computed for a general exponential model but can be computed for a decomposable model.
Decomposable model is an exponential model:

- Represented by a junction tree T.
- Nodes of T = maximal itemsets of $\mathcal{F}$.
- If $a \in X, Y$, then X and Y are connected and every itemset in the path contains a.



Toy junction tree.

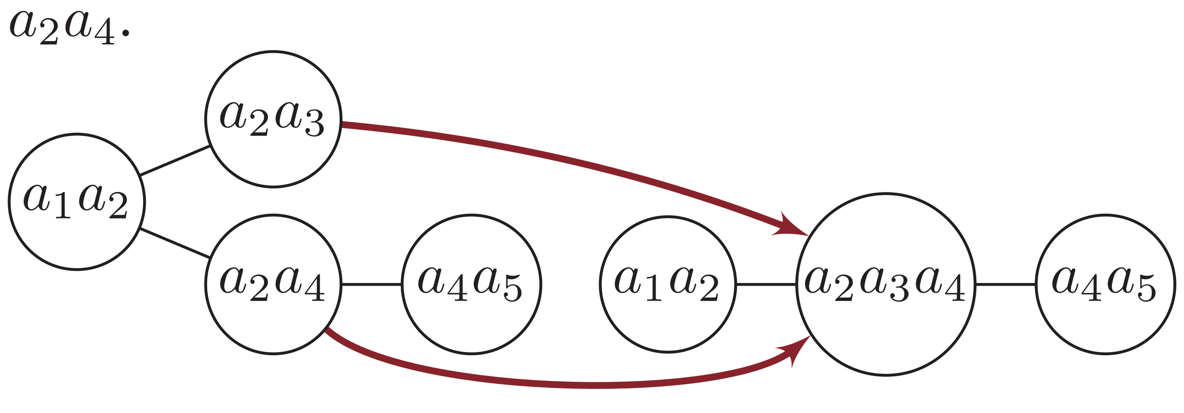
Sampling

Instead of computing the exact score sample N models from $p(M \mid D)$. Estimate the score by

$$sc(X) \approx \frac{\text{number of models containing } X}{N}.$$

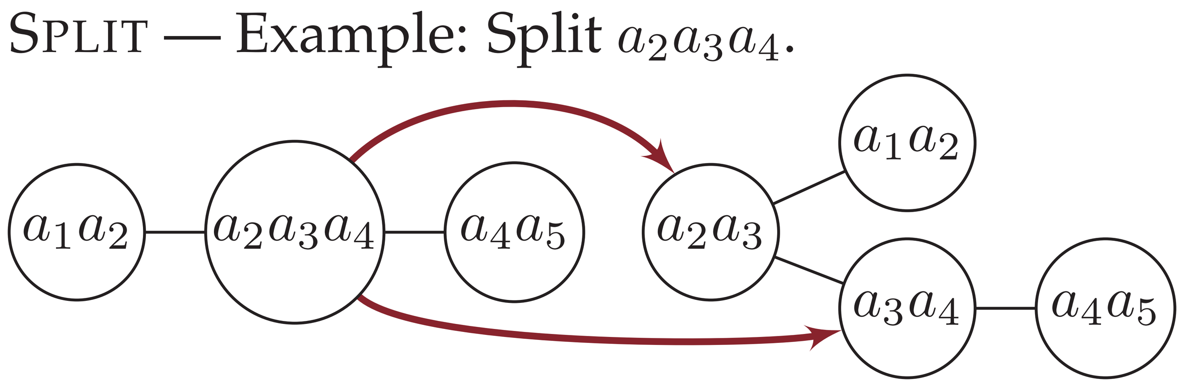
Use MCMC to sample the models. A single MCMC step modifies the junction tree representing the current decomposable model.

Merge — Example: Merge a_2a_3 and a_2a_4.



Before After

Split — Example: Split $a_2a_3a_4$.



Ideal Case

Assume that

- Model M can explain the data.
- $|fam(M)|$ is the smallest among all models that can explain the data.

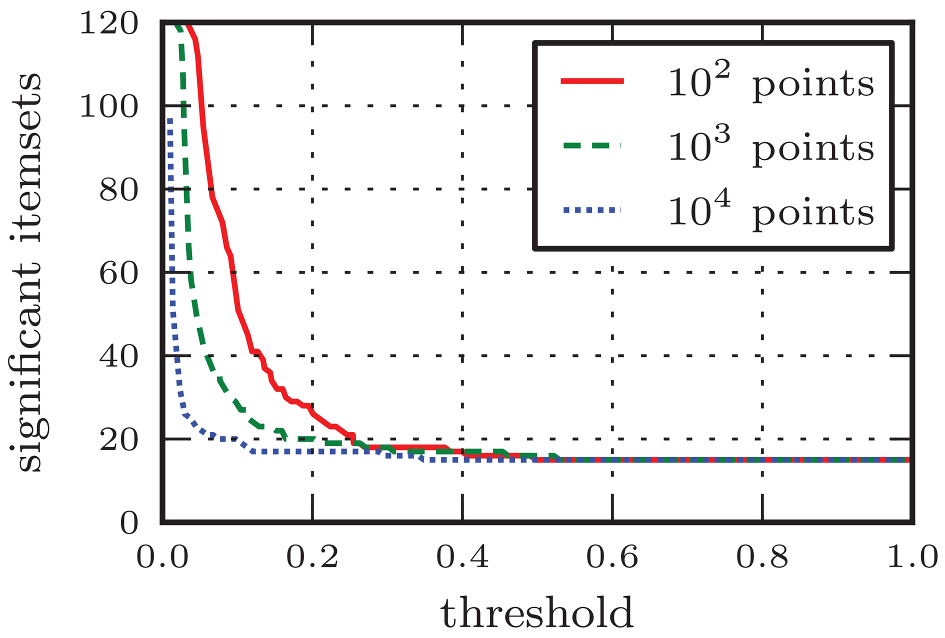
Then, as the number of data points increases,

- $sc(X) \rightarrow 1$, if $X \in fam(M)$,
- $sc(X) \rightarrow 0$, if $X \notin fam(M)$.

Score selects the *minimal* set of itemsets that can explain the data.

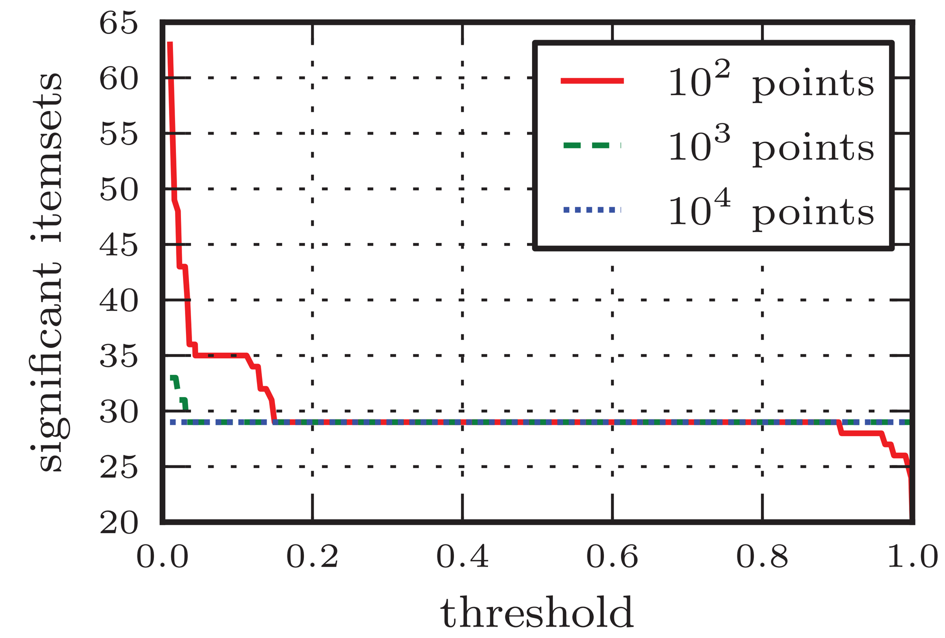
Synthetic Datasets

Synthetic data with 15 independent items. Ideally, $sc(X) = 1$ for singletons and $sc(X) = 0$ for the rest itemsets.



Approaching ideal case: 15 itemsets

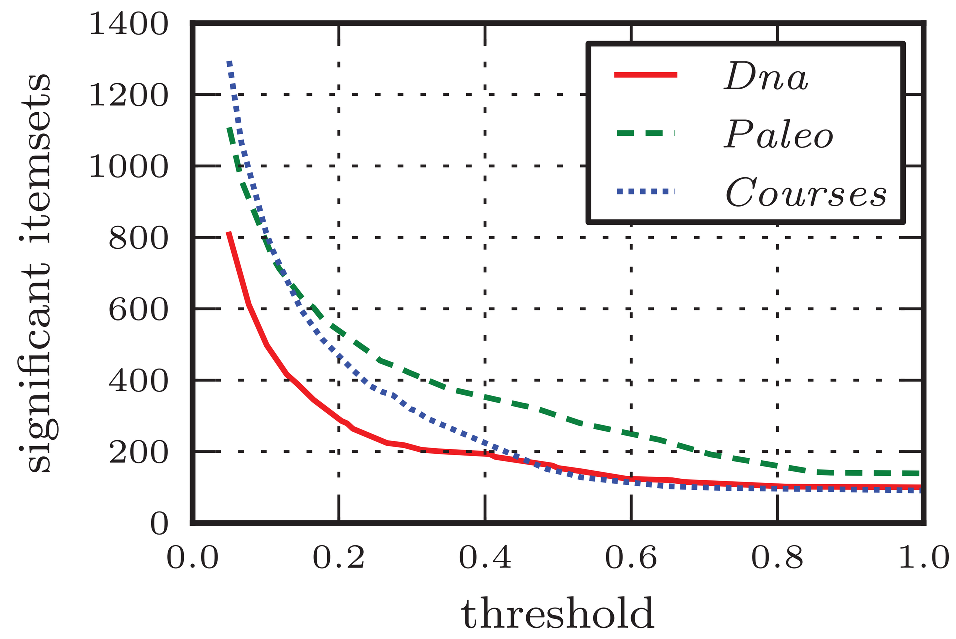
Synthetic data with 15 dependent items, item a_i depends only on a_{i-1}. Ideally, $sc(X) = 1$ for singletons and itemsets $a_{i-1}a_i$, and $sc(X) = 0$ for the rest itemsets.



Approaching ideal case: 29 itemsets

Real-world Datasets

Paleo — species fossils found in specific paleontological sites in Europe.
Courses — enrollment records of students taking courses at the Department of Computer Science of the University of Helsinki.
Dna — DNA copy number amplification data collection of human neoplasms.



Significant itemsets with real-world data