Expect the Unexpected

On the Significance of Subgroups

Matthijs van Leeuwen, Antti Ukkonen

19 Oct 2016

Disclaimer

This talk will probably raise more questions than that it answers.

Subgroup discovery
Problem statement

Given

a database D over attributes A and target T
a pattern language L
a quality measure q
a minimum quality threshold *minqual*

Find each subgroup $s \in L$ for which

s satisfies $q(s) \geq$ *minqual* on D

where $q(s)$ quantifies both:

the **frequency** of s, and
the **deviation** in the distribution of T
for tuples selected by s (compared to D)

Subgroup discovery
Target concept very flexible

Boolean attribute

Numeric attribute

This talk

Exceptional Model Mining

- Regression
- Multi-label
- Networks / graphs
- Preferences
- …

Mining patterns is easy …

… but how do we distinguish
true patterns from **false discoveries**?

Statistics

A common non-parametric test for SD with a numeric target[1]

Test statistic $\theta(s)$
sum of target values of subgroup s

Null hypothesis
$\theta(s)$ is not different from that of a random subset of size $|s|$

Distribution under H_0
$\theta(s)$ for all subsets of size $|s|$

[1] We deal with Boolean targets by considering the proportion of ones.

The common approach
Monte Carlo / permutation sampling

1. S = sample N subsets of size |s| from D

2. Empirical, one-tailed p-value:

$$p(s) = \frac{\#\{H \in S | \theta(H) \geq \theta(s)\}}{N}$$

Multiple hypothesis testing

We test many patterns and need to correct for this

Apply *Bonferroni correction* to control family wise error rate (FWER)

Multiply each p-value with #patterns

Important: count all candidate patterns that were considered during search

Statistical testing for pattern mining

Advantages

+ Principled

+ Can be done post hoc, with any miner

+ Specialised algorithms

Limitations

- Resolution of empirical p-values limited

- Redundancy

- MHT correction

- Assumptions

Exchangeability in subgroup discovery
An essential assumption

All subsets (of a given size) **are equal**

Is this assumption realistic?

Important observation

Pattern mining methods **search** for the 'best' pattern(s) in a language

Not a surprise, that's what they are for

Hence, the top-1 pattern is *not* just any 'random' observation

Correct for **all** candidates (Webb 2007)

In fact, one could skip search and 'test' pattern languages instead

Sample and effect size
Notation and definitions

Sample size *k*

Subgroup size / coverage, i.e., |s|

Effect size

$$q(s) = \frac{\mu(s) - \mu}{\sigma}$$

Subsets and accessibility
Notation and definitions

All subsets

Any subset of D

Accessible subsets $X_{L,D}$

Any subset $E \subseteq D$ for which
$\exists\ s \in L$ s.t. cover(s) = E

Accessibility depends on language L

And data D

Usually $|X_{L,D}| \ll 2^{|D|}$

The odds of finding a large-effect subset

Between accessible and all subsets

We compare the **accessible subsets** to **all** (random) **subsets** (having the same size distribution)

Underlying idea: if accessible subsets have larger effect sizes, then **accessible and all subsets are not exchangeable**

The odds of finding a large-effect subset

Between accessible and all subsets

Formally:

$$\text{odds}(\theta, L, D) = \frac{\Pr\left(q(S) \geq \theta \middle| X_{L,D}\right)}{\Pr(q(S) \geq \theta | D)}$$

where S is any subset in $X_{L,D}$ or from D

Technical details omitted; see paper

Odds of large effect is high
in practical mining settings

Dataset	Target	θ	Odds
Abalone	numeric	1.0	**42.3**
Helsinki	numeric	1.0	**98.9**
Housing	numeric	1.0	**5.7**
Adult	Boolean	0.2	**30.1**
Breast cancer	Boolean	1.0	**9.6**
Spambase	Boolean	1.0	**53.1**

L = all descriptions up to length 3
minimum coverage = 0.05 $|D|$

Expect the Unexpected

Accessible subsets are likely to be significant

First of all: statistical testing is **useful**

- Let's avoid any misunderstanding here

It has its **limitations** though

- Exchangeability of all subsets is too (?) weak
- As witnessed by the high odds

We **empirically** studied this

- And developed the estimators to do this

Open questions

What odds do we obtain on **random data**?

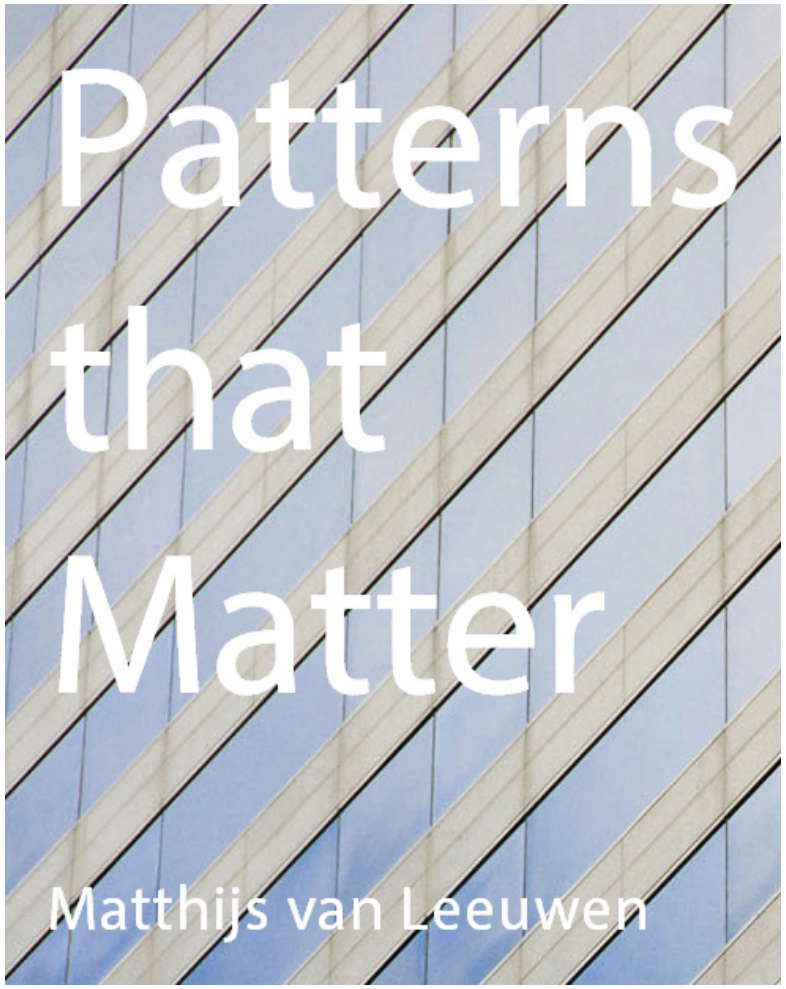
And what does this imply?

How can we easily assess—or even test—**pattern languages**?

Without testing all individual patterns

If exchangeability of all subsets is too weak, then what is a **better assumption**?

All accessible subsets of the same size?



www.patternsthatmatter.org

Antti Ukkonen

www.anttiukkonen.com