Tell Me Something I Don't Know: Randomization Strategies for Iterative Data Mining

Sami Hanhijärvi, Markus Ojala, Niko Vuokko,
Kai Puolamäki, Nikolaj Tatti, Heikki Mannila

HELSINGIN YLIOPISTO
HELSINGFORS UNIVERSITET
UNIVERSITY OF HELSINKI

HELSINKI INSTITUTE FOR INFORMATION TECHNOLOGY

TKK

Motivation

- **Exploratory data mining**
- **Different algorithms used**
  - **Output different types of patterns**
    - **Itemsets, clusters, subgraphs, etc..**
  - **Obtain collections of results**
- **Do all results present different information?**
  - **If they look different, do they present the same information?**

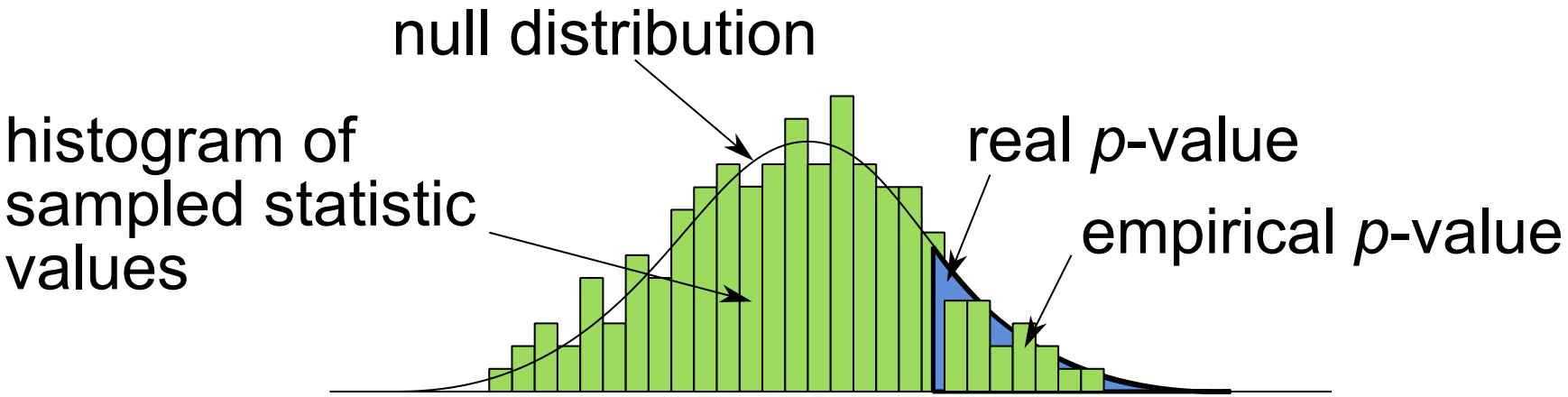
Basic idea

- **Use randomization to assess statistical significance**
- **First tell the user something that is significant**
- **Then tell something that is not implied by previous information**
  - **Randomize while maintaining previous information**
- **Iterate**

Outline of the presentation

- Problem and basic idea
- Motivating example
- Approach and solutions
- Iterative data mining examples
- Conclusions

Empirical p-values



- **Choose a test statistic**
- **Null distribution**
  - **Use randomization**
- ***p*-value**
- **Multiple hypothesis testing**

Motivating example

- **Locations where**
  - **Rat**
  - **Gorilla**
  - **Bird**
  - **Otter**

  have been found
- **Assuming independence, co-occurrence by chance?**

Original dataset

	R	G	B	O
Location 1	1	1	0	0
Location 2	1	1	0	0
Location 3	1	1	1	1
…	1	1	1	1
	0	0	1	1
	0	0	1	1
	0	0	0	0
	0	0	0	0

Motivating example

- **Locations where**
  - **Rat**
  - **Gorilla**
  - **Bird**
  - **Otter**

  have been found
- **Assuming independence, co-occurrence by chance?**

1st randomized

R	G	B	O
1	0	1	1
0	1	0	0
1	1	0	1
0	1	0	0
0	1	1	0
1	0	1	1
0	0	1	0
1	0	0	1

Motivating example

- **Locations where**
  - **Rat**
  - **Gorilla**
  - **Bird**
  - **Otter**

  have been found
- **Assuming independence, co-occurrence by chance?**

2nd randomized

R	G	B	O
0	1	0	0
1	1	0	1
1	0	1	1
1	0	1	0
0	1	0	0
0	0	1	1
0	0	1	0
1	1	0	1

Motivating example

- **Locations where**
  - **Rat**
  - **Gorilla**
  - **Bird**
  - **Otter**

  have been found
- **Assuming independence, co-occurrence by chance?**

3rd randomized

R	G	B	O
1	0	0	1
1	1	1	0
0	0	1	0
0	1	1	1
1	0	0	0
0	1	0	1
0	0	0	1
1	1	1	0

Motivating example

- **Locations where**
  - **Rat**
  - **Gorilla**
  - **Bird**
  - **Otter**

  have been found
- **Assuming independence, co-occurrence by chance?**
- **All co-occurrences significant**

N^{th} randomized

R	G	B	O
1	1	1	0
1	0	1	1
0	0	1	0
1	0	1	0
0	1	0	0
0	0	0	1
1	1	0	1
0	1	0	1

Motivating example

- Fix the information about RG
- Assume independence, given the co-occurrence count of RG
- What else is there in the data?

Original dataset

R	G	B	O
1	1	0	0
1	1	0	0
1	1	1	1
1	1	1	1
0	0	1	1
0	0	1	1
0	0	0	0
0	0	0	0

Motivating example

- Fix the information about RG
- Assume independence, given the co-occurrence count of RG
- What else is there in the data?

1st randomized

R	G	B	O
0	0	1	1
1	1	0	0
0	0	1	0
1	1	0	1
0	0	1	1
1	1	0	0
1	1	0	1
0	0	1	0

Motivating example

- Fix the information about RG
- Assume independence, given the co-occurrence count of RG
- What else is there in the data?

2nd randomized

R	G	B	O
0	0	1	0
0	0	1	0
1	1	1	1
1	1	0	0
0	0	0	1
1	1	0	0
0	0	1	1
1	1	0	1

Motivating example

- **Fix the information about RG**
- **Assume independence, given the co-occurrence count of RG**
- **What else is there in the data?**

3rd randomized

R	G	B	O
1	1	0	0
0	0	1	1
0	0	0	0
1	1	0	1
1	1	1	1
0	0	1	0
0	0	0	1
1	1	1	0

Motivating example

- Fix the information about RG
- Assume independence, given the co-occurrence count of RG
- What else is there in the data?
- RG, RGB and RGO by chance
  - Co-occurrence count of RG explains these
- BO, RBO, GBO and RGBO still significant

N^{th} randomized

R	G	B	O
1	1	0	1
0	0	1	0
1	1	1	1
0	0	0	0
0	0	1	1
1	1	0	0
0	0	0	1
1	1	1	0

Motivating example

- Fix the information about RG and BO
- Assume independence, given co-occurrence counts of RG and BO
- What else is there in the data?

Original dataset

R	G	B	O
1	1	0	0
1	1	0	0
1	1	1	1
1	1	1	1
0	0	1	1
0	0	1	1
0	0	0	0
0	0	0	0

Motivating example

- Fix the information about RG and BO
- Assume independence, given co-occurrence counts of RG and BO
- What else is there in the data?

1st randomized

R	G	B	O
1	1	0	0
0	0	1	1
1	1	1	1
0	0	0	0
0	0	1	1
1	1	0	0
1	1	1	1
0	0	0	0

Motivating example

- Fix the information about RG and BO
- Assume independence, given co-occurrence counts of RG and BO
- What else is there in the data?

2nd randomized

R	G	B	O
0	0	1	1
1	1	0	0
0	0	0	0
1	1	1	1
1	1	0	0
0	0	1	1
0	0	1	1
1	1	0	0

Motivating example

- Fix the information about RG and BO
- Assume independence, given co-occurrence counts of RG and BO
- What else is there in the data?

3rd randomized

R	G	B	O
0	0	0	0
0	0	1	1
1	1	0	0
1	1	0	0
0	0	0	0
0	0	1	1
1	1	1	1
1	1	1	1

Motivating example

- Fix the information about RG and BO
- Assume independence, given co-occurrence counts of RG and BO
- What else is there in the data?
- All else by chance
- Conclusion: Column sums and RG and BO explain data

N^{th} randomized

R	G	B	O
1	1	0	0
0	0	1	1
0	0	0	0
1	1	1	1
1	1	0	0
0	0	0	0
1	1	1	1
0	0	1	1

Problem

- **Do different results explain same the structure?**
- **How to iteratively tell something new?**
- **We consider only 0-1 data**

Outline of the presentation

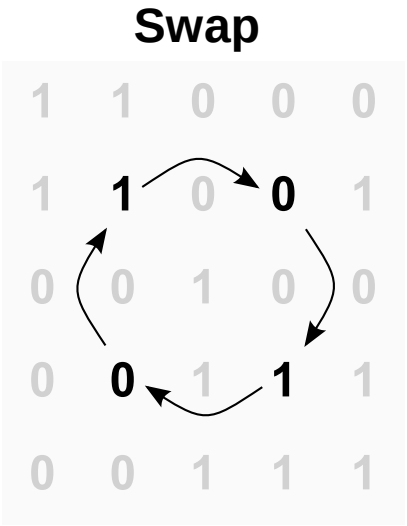
- Problem and basic idea
- Motivating example
- **Approach and solutions**
- Iterative data mining examples
- Conclusions

Approach

- **Use randomization to obtain empirical *p*-values**
- **First assess if results significant in general case**
- **Iterate**
  - **Select part of significant result**
  - **Constraint randomization with that part**
  - **Assess the rest**
- **All information of non-significant results contained in constraints**

Solutions

- **Baseline**
  - Maintain column and row sums (margins)
  - Swap randomization (Cobb *et al.* 2003)
- **Randomize and maintain frequencies of selected itemsets**
  - Exact solution NP-hard
- **Maintain frequencies approximately**
  - Swap randomization is a Markov-chain
  - Use Metropolis-Hastings

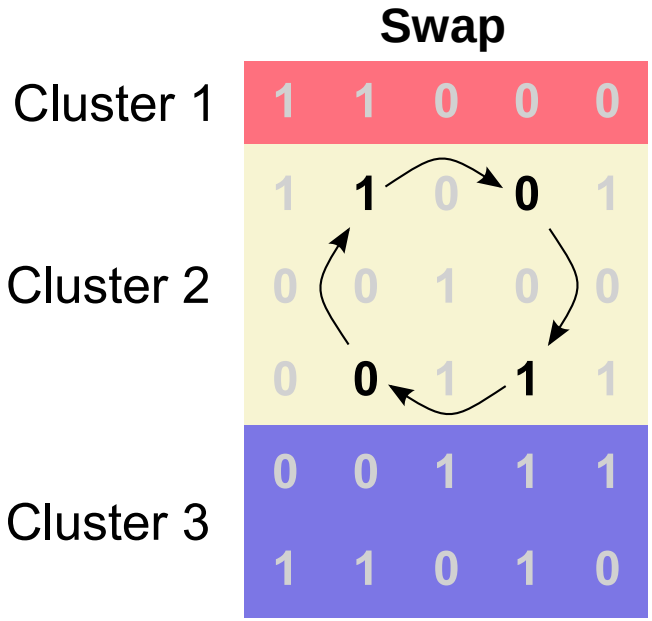


Solutions

- **Iterate (standard Metropolis-Hastings)**
  - Find a random swap
  - Calculate error in maintained frequencies
  - Perform swap
    - Always, if error does not increase
    - With decreasing probability in increasing error
- **With enough swaps, matrix is random while satisfying constraints**
  - Mixing time hard to estimate
  - Besag and Clifford

Solutions

- **Randomize and maintain cluster structure**
  - **Clustering on rows**
  - **Randomize in each cluster separately**
  - **Keeps clustering error constant**



TKK

Outline of the presentation

- Problem and basic idea
- Motivating example
- Approach and solutions
- **Iterative data mining examples**
- Conclusions

Iterative data mining examples

- **Significant itemsets of size 2**
  - **Compare itemset and clustering constraints**
- **Discovering significant itemsets iteratively**
  - **Example iterative process**

Significant itemsets of size 2

- **Assess the significance of all possible size 2 itemsets**
- **Paleo dataset**
  - **Items are species**
  - **Transactions are excavation sites**

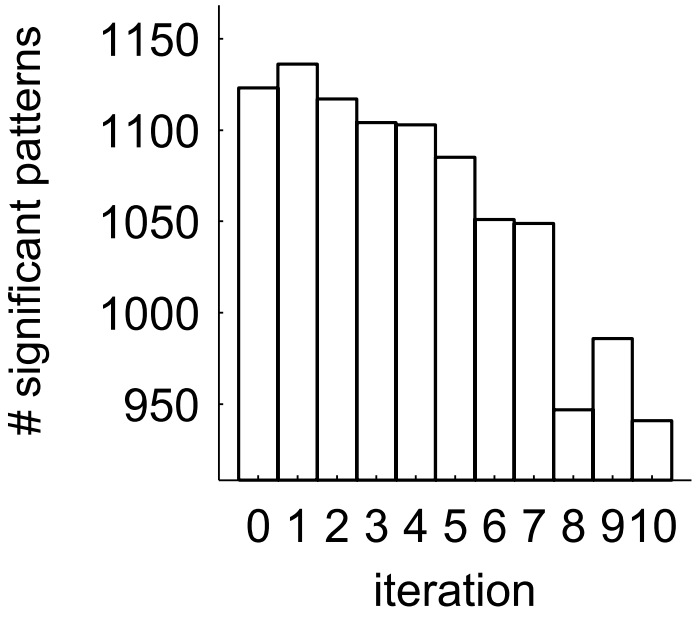
Significant itemsets of size 2

- **Randomize and maintain clustering structure**
  - **With 3 clusters, no itemset significant**
- **Randomize and maintain frequencies of adjacent itemsets**
  - **Some itemsets significant**
  - **Clustering significant**

Discovering significant itemsets iteratively

- **Only consider itemsets of size 2 and 3**
  - **Count number of significant itemsets**
- **Iteratively add one itemset to constraints**
  - **Select itemset with smallest *p*-value**
  - **Count number of significant itemsets**

Discovering significant itemsets iteratively



Iterative data mining examples

- **More detailed experiments in the paper and poster**

Paleo		CM	
		N	S
IM	N	8977	**0**
	S	614	**0**

Paleo		M	
		N	S
IM	N	**8832**	145
	S	60	**544**

Outline of the presentation

- Problem and basic idea
- Motivating example
- Approach and solutions
- Iterative data mining examples
- **Conclusions**

Conclusions

- **Different results may explain the same phenomenon**
- **Iteratively find significant information in data**
- **Statistical significance to measure overlap**
- **Example shown in specific problem area**

- **Questions?**
- **Comments?**
- **More information at the poster in today's session**