From Black and White to Full Colour

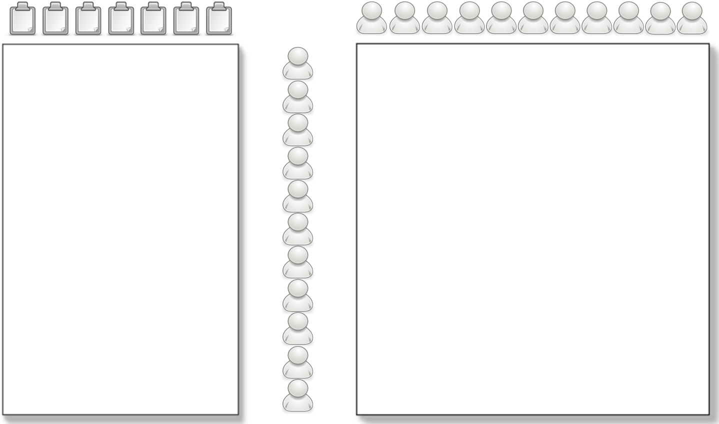
Extending Redescription Mining Outside the Boolean World

Esther Galbrun   Pauli Miettinen

Helsinki Institute for Information Technology
Department of Computer Science, University of Helsinki

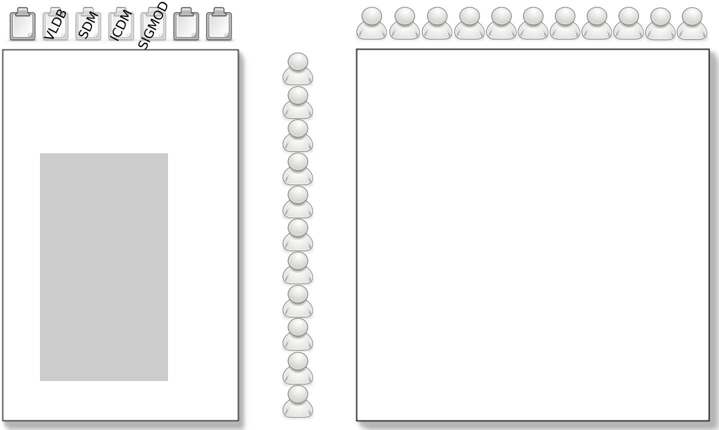
Max-Planck Institute for Informatics

Example: DBLP Data



Example: DBLP Data

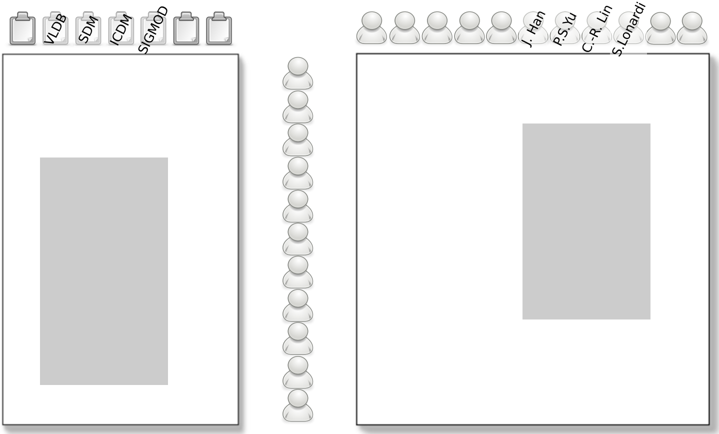
VLDB ∧ ICDM ∧ SDM ∧ SIGMOD



Example: DBLP Data

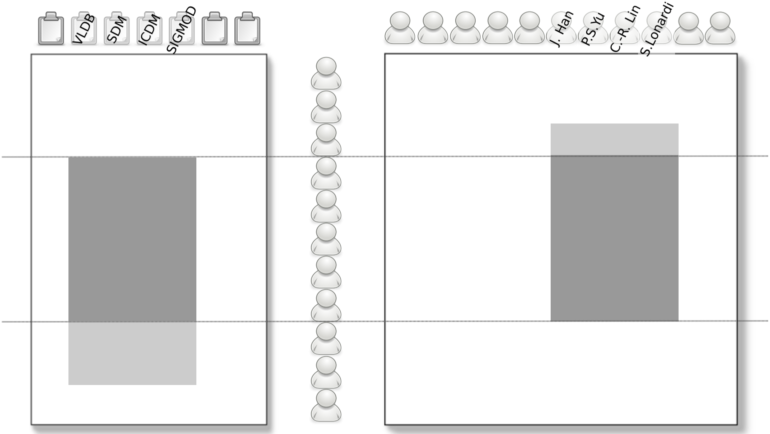
VLDB ∧ ICDM ∧ SDM ∧ SIGMOD

(J. Han ∧ P.S. Yu) ∨ C.-R. Lin ∨ S. Lonardi



Example: DBLP Data

VLDB ∧ ICDM ∧ SDM ∧ SIGMOD
(J. Han ∧ P.S. Yu) ∨ C.-R. Lin ∨ S. Lonardi

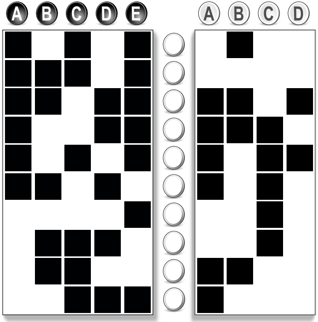


Definitions

Redescription Given two datasets with identity between the rows, a **redescription** is a pair of queries $(q_{\mathbf{L}}, q_{\mathbf{R}})$ over the columns characterizing approximately the same sets of rows.

Redescription Mining Given such a pair of datasets and a set of constraints, find the best redescriptions satisfying the constraints.

Definitions

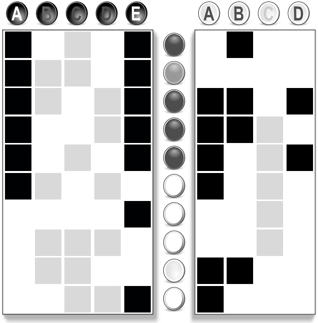


Dataset Boolean matrices

Definitions

$\boxed{A} \wedge \boxed{E}$

$(A \wedge D) \vee B$



Dataset Boolean matrices

Queries Boolean formulae

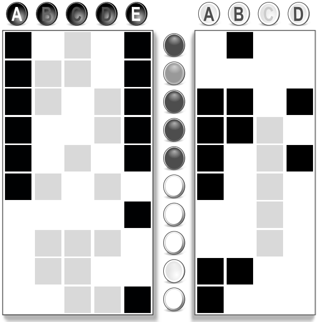
Accuracy Jaccard coefficient

$$J(q_{\mathbf{L}}, q_{\mathbf{R}}) = \frac{|\mathrm{supp}(q_{\mathbf{L}}) \cap \mathrm{supp}(q_{\mathbf{R}})|}{|\mathrm{supp}(q_{\mathbf{L}}) \cup \mathrm{supp}(q_{\mathbf{R}})|} = \frac{\left|\boxed{E_{1,1}}\right|}{\left|\boxed{E_{1,0}}\right| + \left|\boxed{E_{1,1}}\right| + \left|\boxed{E_{0,1}}\right|}$$

Definitions

$\boxed{A} \wedge \boxed{E}$

$(A \wedge D) \vee B$

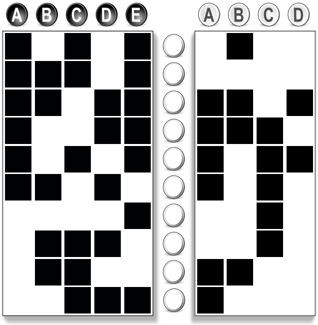


Dataset	Boolean matrices
Queries	Boolean formulae
Accuracy	Jaccard coefficient
Constraints	Support, accuracy, length of the query, *p*-value, …

Special Cases

$$\boxed{?} \wedge \boxed{?} \Rightarrow ? \wedge ? \wedge ?$$

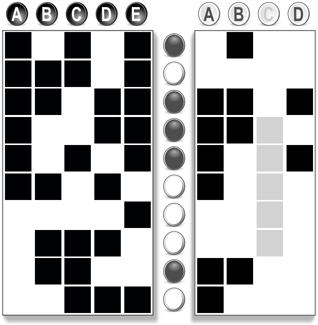
$$\boxed{?} \wedge \boxed{?} \Leftarrow ? \wedge ? \wedge ?$$



Only conjunctive queries:
bi-directional
association rules

Special Cases

$\boxed{?}\ ?\ \boxed{?} \Rightarrow (A \wedge D) \vee B$



One query given: classification task

Algorithms for Redescription Mining

Different approaches to Boolean Redescription Mining:

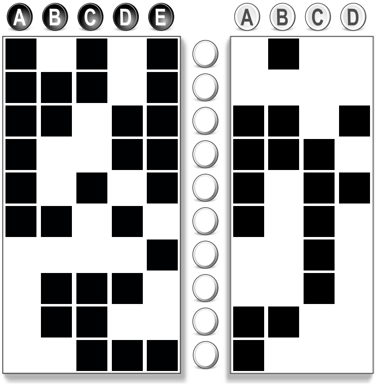
Decision trees: Ramakrishnan et al. 2004 (`CARTwheels`)

Co-clusters: Parida and Ramakrishnan, 2005

Frequent Itemsets: Gallo, Miettinen and Mannila, 2008

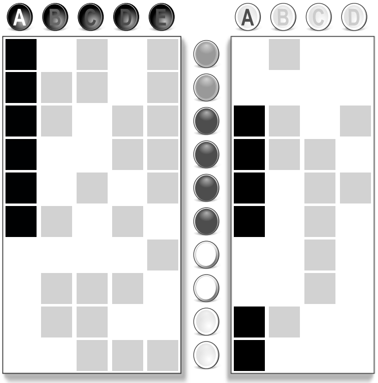
Greedy: *Eidem*

Greedy Query Extension



Greedy Query Extension

Initial pair

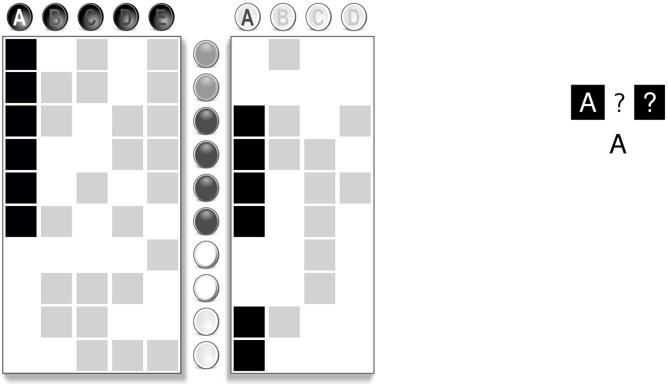


A

A

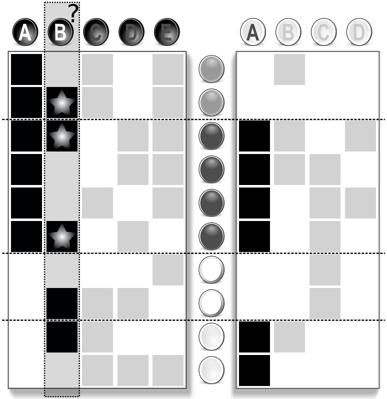
Greedy Query Extension

Try to extend one side



Greedy Query Extension

Try to append ∧ B to left hand side

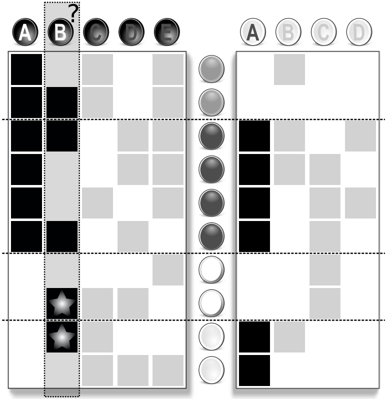


A ∧ B

A

Greedy Query Extension

Try to append $\vee$ B to left hand side

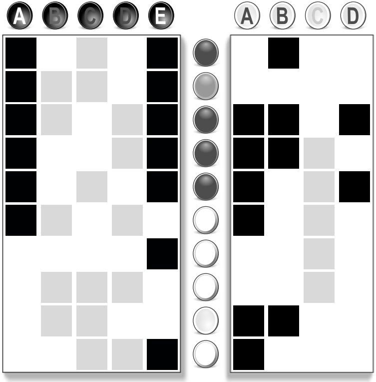


$\boxed{A} \vee \boxed{B}$

A

Greedy Query Extension

…after few iterations



$$\boxed{A} \wedge \boxed{E}$$

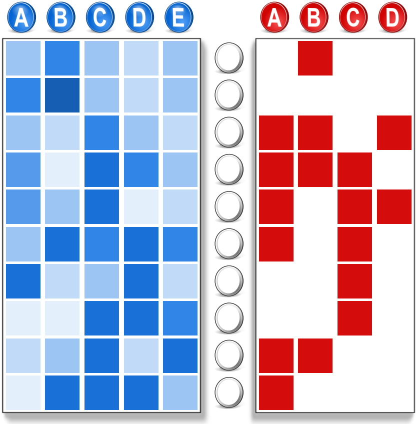
$$(A \wedge D) \vee B$$

$$= \frac{4}{1 + 4 + 1}$$

$$= 0.66$$

What if your data is not Boolean?

What if your data is not Boolean?



What if your data is not Boolean?

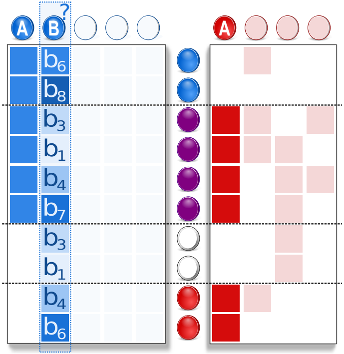
Existing methods Discretization as a pre-processing step

- Use one variable per category
- Bucketing real-valued attributes
- Explosion of the number of variables
- Requires extensive domain knowledge

Our approach Discretization within the algorithm

- Optimal interval determined on-the-fly
- No pre-processing

Greedy Query Extension



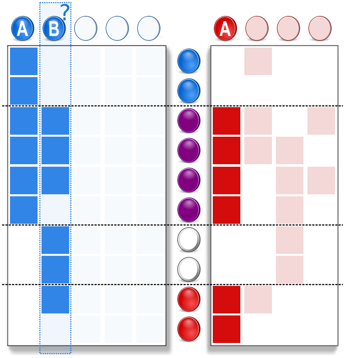
$$[\, a_\lambda \le A \le a_\rho \,] \;\wedge\; [\lambda \le B \le \rho]$$

A

$$J(q_{\mathbf{L}} \wedge [\lambda \le B \le \rho], q_{\mathbf{R}})$$

$$= \frac{|E_{1,1}([\lambda \le B \le \rho])|}{|E_{1,0}([\lambda \le B \le \rho])| + |E_{0,1}| + |E_{1,1}|}$$

Greedy Query Extension



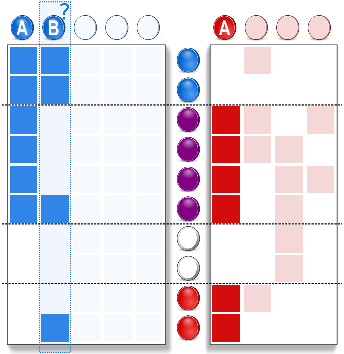
$$[\, a_\lambda \leq A \leq a_\rho \,] \;\wedge\; [\, b_1 \leq B \leq b_4 \,]$$

A

$$\mathrm{J}(q_{\mathbf{L}} \wedge [\lambda \leq B \leq \rho], q_{\mathbf{R}})$$

$$= \frac{|E_{1,1}([\lambda \leq B \leq \rho])|}{|E_{1,0}([\lambda \leq B \leq \rho])| + |E_{0,1}| + |E_{1,1}|}$$

Greedy Query Extension

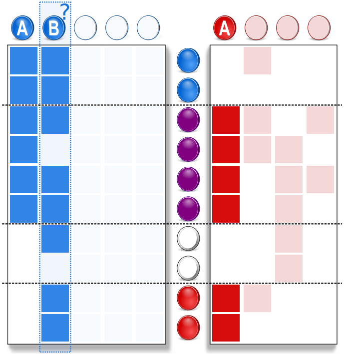


$$[a_\lambda \leq A \leq a_\rho] \land [b_5 \leq B \leq b_9]$$

A

$$J(q_L \land [\lambda \leq B \leq \rho], q_R) = \frac{|E_{1,1}([\lambda \leq B \leq \rho])|}{|E_{1,0}([\lambda \leq B \leq \rho])| + |E_{0,1}| + |E_{1,1}|}$$

Greedy Query Extension

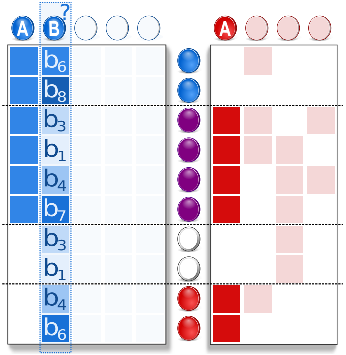


$$[\, a_\lambda \leq A \leq a_\rho \,] \;\underset{A}{\wedge}\; [\, b_2 \leq B \leq b_8 \,]$$

$$J(q_\mathbf{L} \wedge [\lambda \leq B \leq \rho], q_\mathbf{R})$$

$$= \frac{\left|E_{1,1}([\lambda \leq B \leq \rho])\right|}{\left|E_{1,0}([\lambda \leq B \leq \rho])\right| + \left|E_{0,1}\right| + \left|E_{1,1}\right|}$$

Greedy Query Extension



$$[\boxed{a_\lambda} \leq A \leq \boxed{a_\rho}] \wedge [\lambda \leq B \leq \rho]$$

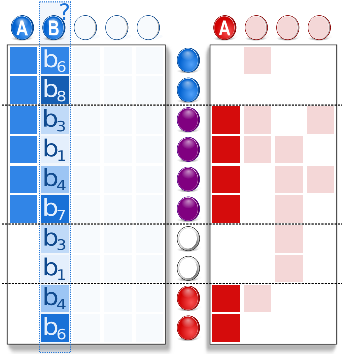
A

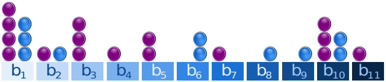
$$J(q_{\mathbf{L}} \wedge [\lambda \leq B \leq \rho], q_{\mathbf{R}})$$

$$= \frac{|E_{1,1}([\lambda \leq B \leq \rho])|}{|E_{1,0}([\lambda \leq B \leq \rho])| + |E_{0,1}| + |E_{1,1}|}$$

Finding the Best Interval

Ordering the values and finding the best cut points



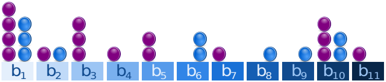


$$J(q_{\mathbf{L}} \wedge [\lambda \leq B \leq \rho], q_{\mathbf{R}})$$

$$= \frac{\left|E_{1,1}([\lambda \leq B \leq \rho])\right|}{\left|E_{1,0}([\lambda \leq B \leq \rho])\right| + \left|E_{0,1}\right| + \left|E_{1,1}\right|}$$

Finding the Best Interval

Proposition:
The optimal value for λ is one of the lower cut points or $-\infty$;
the optimal value for ρ is one of the upper cut points or $+\infty$.



$$\frac{\left|E_{1,1}([\lambda \leq B \leq \rho])\right|}{\left|E_{1,0}([\lambda \leq B \leq \rho])\right| + \left|E_{0,1}\right| + \left|E_{1,1}\right|}$$

A **lower cut point** is a value b_i such that $b_i \in D(E_{1,1}, B)$ and $b_{i-1} \in D(E_{1,0}, B)$.
An **upper cut point** is a value b_j such that $b_j \in D(E_{1,1}, B)$ and $b_{j+1} \in D(E_{1,0}, B)$.

Similar to result for classification learning by Fayyad and Irani, 1993.

The ReReMi algorithm

```
for each best singleton redescription do
    while there are extendable redescriptions do
        try to extend the queries
        for each free variable and each Boolean operator do
            find the best interval
        select the best extensions
return the redescriptions
```

Synthetic data

- Generate pairs of matrices containing a redescription, Boolean/Boolean vs. Boolean/Real-Valued
- Add noise, conservative vs. destructive

✓ The algorithm finds the planted redescriptions, except in a handful of cases where they do not comply with mining constraints

Comparison to association rule mining

- Mine for frequent itemsets using off-the-shelf tool,
  construct bi-directionnal association rules
- Mine redescriptions with our algorithm,
  only conjunctions of positive literals

✓ Found the strongest rules, much less redundancy

Comparison with `CARTwheels`

$\mathtt{DBLP_B}$ data

q_L	q_R	J
`CARTwheels`		
(STOC $\wedge \neg$FOCS) $\vee \neg$STOC	B. Dageville $\vee$ ($\neg$B. Dageville $\wedge \neg$A. Wigderson)	0.736
ICDM $\vee$ ($\neg$ICDM $\wedge \neg$STOC)	(C. Olston $\wedge \neg$C. Chekuri) $\vee$ ($\neg$C. Olston $\wedge \neg$A. Wigderson)	0.691
`ReReMi`		
STOC $\wedge$ COLT $\wedge$ ICML	Y. Freund $\vee$ N. Littlestone $\vee$ P.M. Long $\vee$ S. Kwek	0.500
ICDM $\wedge$ SDM $\wedge$ KDD	J. Lin $\vee$ I.S. Dhillon $\vee$ P.S. Yu $\vee$ V. Kumar	0.338

✓ Many negations, quick drop in accuracy vs. easy to interpret, zero *p*-values

Pre-bucketing

- Comparing preprocessing to on-the-fly bucketing
- Boolean/real-valued dataset
- Discretize the data
  - Methods: equal width, equal height, segmentation
  - Number of bins: from 10 to 150
- Mine with `CARTwheels` and Boolean `ReReMi`
- Select the discretization with best results
- Compare to real-valued `ReReMi`

✓ On-the-fly bucketing yielded best results

Application to Bioclimatic Niche Finding

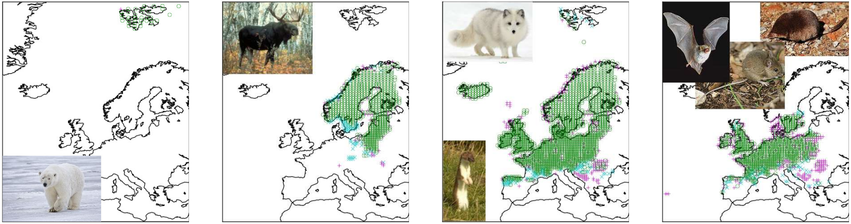
Dataset: Spatial land areas of Europe (2575 entities)

- Presence/absence of mammals
  (194 species)
- Climatic data
  (48 temperature and rainfall variables)

Question: Find a query over climatic variables that describes the area inhabited by (a group of) mammal species (and vice versa)

Application to Bioclimatic Niche Finding

q_L	q_R	J	supp
(1) Polar Bear	$[-7.0727 \leq t^{avg}_{May} \leq -3.375]$	0.973	36
(2) European Elk	$([-9.80 \leq t^{max}_{Fev} \leq 0.40] \wedge [12.20 \leq t^{max}_{Jul} \leq 24.60]$ $\wedge [56.852 \leq p^{avg}_{Aug} \leq 136.46]) \vee [183.27 \leq p^{avg}_{Sep} \leq 238.78]$	0.814	582
(3) Arctic Fox $\vee$ Stoat	$(([2.60 \leq t^{max}_{Jun} \leq 8.50] \vee [7.20 \leq t^{max}_{Sep} \leq 22.20])$ $\wedge [36.667 \leq p^{avg}_{Aug}]) \vee [21.133 \leq t^{avg}_{Jul} \leq 21.20]$	0.813	1477
(4) Wood Mouse $\wedge$ Natterer's Bat $\wedge$ Eurasian Pygmy Shrew	$([3.20 \leq t^{max}_{Mar} \leq 14.50] \wedge [17.30 \leq t^{max}_{Aug} \leq 25.20]$ $\wedge [14.90 \leq t^{max}_{Sep} \leq 22.80]) \vee [19.60 \leq t^{avg}_{Jul} \leq 19.956]$	0.623	681



t^{min}_X, t^{max}_X, and t^{avg}_X stand for minimum, maximum, and average temperature of month X in Celsius. p^{avg}_X stands for average precipitation of month X in *mm*. Green circles, cyan plus signs and magenta crosses respectively indicate areas where both queries hold, only the left query holds and only the right query holds.

Future work

- Applications:
  - Niche-finding using **trait** data
  - Other domains, e.g. medical data
- Improve the algorithm, e.g. computation of initial pairs
- Proofs of the behavior of the algorithm

Conclusions

- Redescription Mining:
  - Interesting and powerful data-mining tool
  - Even more powerful extended to real-valued data
- On-the-fly bucketing approach:
  - Fast, better than existing bucketing approaches
  - Possibly applicable to other data-mining problems

Implementation available online:

`http://www.cs.helsinki.fi/u/galbrun/redescriptors/`

Conclusions

- Redescription Mining:
  - Interesting and powerful data-mining tool
  - Even more powerful extended to real-valued data
- On-the-fly bucketing approach:
  - Fast, better than existing bucketing approaches
  - Possibly applicable to other data-mining problems

Implementation available online:

`http://www.cs.helsinki.fi/u/galbrun/redescriptors/`

Thank you …

...Questions?

Queries

The queries are boolean formulae, i.e. literals and their negations connected with logical conjunction ($\wedge$) and disjunction ($\vee$).

- Every variable appears only once
- No operator precedence, queries can be parsed in linear order without trees

✗ $a \vee \neg a$

✗ $(a \wedge b) \vee (c \wedge d)$

✓ $(a \vee b) \wedge \neg c$

- Expressivity vs. interpretability
- Reasonable search space

Constraints

Accuracy: Jaccard coefficient

Support: number of entities covered

Contribution: number of entities contributed by a variable

Length: number of variables in the queries

p-value: statistical significance of the queries

Type of query: conjunctions, disjunctions, negations

Applying the support constraints

- Support monotonicity does not hold
- Use (softer) constraints to prune the search space
- Filter the end results
- Faster search vs.
  risk of discarding potentially good candidates

Statistical significance

Different null hypotheses

Redescription: independent queries

$$\mathrm{pvalM}(q_{\mathbf{L}}, q_{\mathbf{R}}) = \sum_{s=|\mathrm{supp}(q_{\mathbf{L}}, q_{\mathbf{R}})|}^{|E|} \binom{|E|}{s} (p_R)^s (1-p_R)^{|E|-s},$$

Conjunctive extension: uncorrelation

$$\mathrm{pvalE}(q_s, \wedge l) = \mathrm{pvalM}(q_s, l)$$

Disjunctive extension: correlation

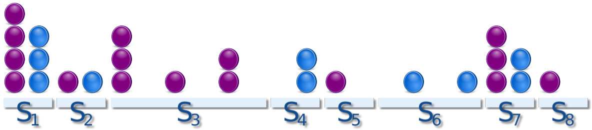
$$\mathrm{pvalE}(q_s, \vee l) = 1 - \mathrm{pvalM}(q_s, l)$$

Interval approximation

When evaluating the accuracies of consecutive intervals S_{i-}, S_i and S_{i+}.

If S_i alone is better than merging S_{i-} and S_i, then S_i alone or merging S_i and S_{i+} is better than S_{i-}, S_i, and S_{i+} for any interval S_{i+}, so S_{i-} can be dropped.

On the other hand, if merging S_{i-} and S_i is better than S_i alone, there might still be an interval S_{i+} such that merging S_i and S_{i+} is better than S_{i-}, S_i and S_{i+} together.



Interval approximation

Let l, i, and u be any indices such that $S_{i-} = [t_l, t_i]$, $S_i = [t_i, t_{i+1}]$, and $S_{i+} = [t_{i+1}, t_u]$ are valid intervals.

$$j(t_l, t_{i+1}) < j(t_i, t_{i+1}) \Rightarrow j(t_l, t_u) < \max(j(t_i, t_{i+1}), j(t_i, t_u))$$

We use this property to find the best interval by upward aggregation.
On the other hand,

$$j(t_i, t_{i+1}) < j(t_l, t_{i+1}) \not\Rightarrow j(t_i, t_u) < j(t_l, t_u)$$

Therefore, we also compute the best interval using downward aggregation and combine the two.

Interval approximation

We can compute an interval that approximates the optimal accuracy in time linear to the number of cut points.

Especially useful when the rows in $E_{1,0}$ and $E_{1,1}$ are not clearly separated, saves heavy computations for variables that are intuitively poor extensions.

Initial pairs

Boolean/Real-Valued

- Consider each of the Boolean variables v in turn and redescription $(v, \emptyset)$
- Find best right hand side disjunctive extension for each real-valued variable

Real-Valued/Real-Valued

- Brute-force search, using interval approximation
- Too expensive for dense data with wide range of values

Comparison with CARTwheels

DBLP$_B$ data

q_L	q_R	J	supp	p-value
CARTwheels				
(STOC $\wedge\neg$FOCS) $\vee\neg$STOC	B. Dageville $\vee$ ($\neg$B. Dageville $\wedge\neg$A. Wigderson)	0.736	1673	0.011
ICDM $\vee$ ($\neg$ICDM $\wedge\neg$STOC)	(C. Olston $\wedge\neg$C. Chekuri) $\vee$ ($\neg$C. Olston $\wedge\neg$A. Wigderson)	0.691	1570	0.017
ReReMi				
STOC $\wedge$ COLT $\wedge$ ICML	Y. Freund $\vee$ N. Littlestone $\vee$ P.M. Long $\vee$ S. Kwek	0.500	21	0.000
ICDM $\wedge$ SDM $\wedge$ KDD	J. Lin $\vee$ I.S. Dhillon $\vee$ P.S. Yu $\vee$ V. Kumar	0.338	44	0.000

Comparison with `CARTwheels`

`CARTwheels`: Many negations, quick drop in accuracy

`ReReMi`: Easy to interpret, zero p-values

Pre-bucketing

- Comparing preprocessing to on-the-fly bucketing
- Boolean/real-valued dataset
- Discretize the data
  - Methods: equal width, equal height, segmentation
  - Number of bins: from 10 to 150
- Mine with `CARTwheels` and Boolean `ReReMi`
- Select the discretization with best results
- Compare to real-valued `ReReMi`

Pre-bucketing

$q_{\mathbf{L}}$, $q_{\mathbf{R}}$	J	supp	p-value
`CARTwheels`			
(European Pine Vole $\wedge$ European Pine Marten) $\vee$ ($\neg$ European Pine Vole),			
$([57.5 \leq p^{\text{avg}}_{\text{Nov}} \leq 62.706] \wedge \neg[75.03 \leq p^{\text{avg}}_{\text{Jun}} \leq 82.6]) \vee (\neg[57.5 \leq p^{\text{avg}}_{\text{Nov}} \leq 62.706])$	0.980	1244	0.007
Boolean `ReReMi`			
(Striped Field Mouse $\vee$ House mouse)$\wedge$ Wood mouse,			
$[5.925 \leq t^{\text{avg}}_{\text{Apr}} \leq 7.0] \vee [7.0 \leq t^{\text{avg}}_{\text{Apr}} \leq 7.9077] \vee [7.9077 \leq t^{\text{avg}}_{\text{Apr}} \leq 8.46] \vee [8.46 \leq t^{\text{avg}}_{\text{Apr}}]$	0.807	442	0.000
`ReReMi`			
Striped Field Mouse $\vee$ European Hedgehog,			
$([6.45 \leq t^{\text{avg}}_{\text{May}}] \wedge [9.3067 \leq t^{\text{avg}}_{\text{Jun}}]) \vee [-11.994 \leq t^{\text{avg}}_{\text{Jan}} \leq -11.888] \vee [192.5 \leq p^{\text{avg}}_{\text{Sep}}]$	0.909	903	0.000

Pre-bucketing

`CARTwheels`: less interesting and quick drop in accuracy

Boolean `ReReMi`: lower accuracies

✓ On-the-fly bucketing yielded best results

Bioclimatic Niche Finding

- A.k.a. bioclimatic envelope finding
- Well-known task for biologists
- Several definitions for the term **niche**
- Restricted to single, hand-selected species
- Methods used include regression, neural networks, genetic algorithms, …

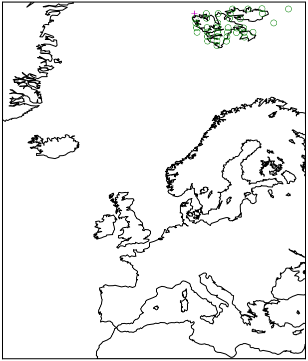
Redescription Mining for Niche Finding

- Automated niche finding
- Allow for more complex sets of species
- Easy-to-understand method
- Possibly generalizable from species to **traits**

Redescription Mining for Niche Finding

Polar Bear

$$[-7.0727 \leq t^{\text{avg}}_{\text{May}} \leq -3.375]$$

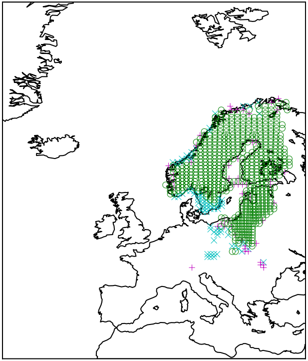


$\text{J} = 0.973 \quad \text{supp} = 36$

Redescription Mining for Niche Finding

European Elk

$$([-9.80 \leq t^{\max}_{\text{Fev}} \leq 0.40] \wedge [12.20 \leq t^{\max}_{\text{Jul}} \leq 24.60] \wedge$$
$$[56.852 \leq p^{\text{avg}}_{\text{Aug}} \leq 136.46]) \vee [183.27 \leq p^{\text{avg}}_{\text{Sep}} \leq 238.78]$$

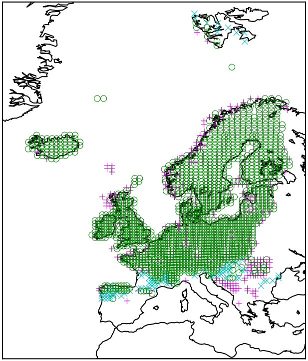


$J = 0.814 \quad \text{supp} = 582$

Redescription Mining for Niche Finding

Arctic Fox $\vee$ Stoat

$$(([2.60 \le t^{\max}_{\mathrm{Jun}} \le 8.50] \vee [7.20 \le t^{\max}_{\mathrm{Sep}} \le 22.20]) \wedge$$
$$[36.667 \le p^{\mathrm{avg}}_{\mathrm{Aug}}]) \vee [21.133 \le t^{\mathrm{avg}}_{\mathrm{Jul}} \le 21.20]$$

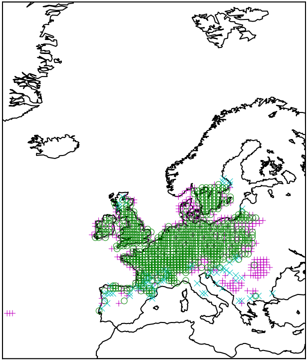


J = 0.813   supp = 1477

Redescription Mining for Niche Finding

Wood Mouse $\wedge$ Natterer's Bat $\wedge$ Eurasian Pygmy Shrew

$$([3.20 \le t^{\max}_{\text{Mar}} \le 14.50] \wedge [17.30 \le t^{\max}_{\text{Aug}} \le 25.20] \wedge$$
$$[14.90 \le t^{\max}_{\text{Sep}} \le 22.80]) \vee [19.60 \le t^{\text{avg}}_{\text{Jul}} \le 19.956]$$



$J = 0.623 \quad \text{supp} = 681$

Esther Galbrun[1 2]
esther.galbrun@cs.helsinki.fi

Pauli Miettinen[3]
pmiettin@mpi-inf.mpg.de

Part of this work was done when the author was with HIIT.

Department of Computer Science
University of Helsinki
Finland

Helsinki Institute for Information Technology
Helsinki
Finland

Max-Planck Institute for Informatics
Saarbrücken
Germany

Illustrations credits:

Natterer's Bat: The Norfolk Bat Group
`http://www.norfolk-bat-group.org.uk`

Artic Fox: Flickr `http://www.flickr.com`

Yawning Polar Bear: LIFE Magazine
`http://www.life.com`

Others: Wikipedia
`http://en.wikipedia.org`