An Introduction to Redescription Mining

Esther Galbrun
esther.galbrun@inria.fr

Young research scientist (CR2)
Inria Nancy – Grand Est
Team Orpailleur
Traditional Data Mining
Traditional Data Mining
Traditional Data Mining

Wildcat ∧ Stoat ∧ Rabbit

31UFP2
32VLK1
33WXR1
35TLN3
Redescription Mining
Redescription Mining

Elk ∨ Alpine Field Mouse

\([-5.3 \leq t^{+}_{\text{Nov}} \leq 6.0] \lor [12.811 \leq t^{\sim}_{\text{Sep}} \leq 12.825]) \land [14.0 \leq t^{+}_{\text{Jul}} \leq 24.4]\)
Redescription Mining

Elk $\lor$ Alpine Field Mouse

$$([−5.3 \leq t^+_{\text{Nov}} \leq 6.0] \lor [12.811 \leq t_{\sim\text{Sep}} \leq 12.825]) \land [14.0 \leq t^+_{\text{Jul}} \leq 24.4]$$
Redescription Mining

Finding different descriptions characterizing a group of entities…
Redescription Mining

...finding entities having different common descriptions.
Definition: Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.
How do we find redescriptions?

Definition: Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.

Particularity: Finding patterns across several sets of attributes.
Definition: Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.

Dataset  Two data matrices
Definition: Given a collection of entities characterized by two sets of attributes, a **redescription** is a pair of queries \((q_L, q_R)\) describing the same set of entities.
How do we find redescriptions?

**Definition:** Given a collection of entities characterized by two sets of attributes, a *redescription* is a pair of queries \((q_L, q_R)\) describing the same set of entities.

**Accuracy** Jaccard coefficient of the supports

\[
J(q_L, q_R) = \frac{|\text{supp}(q_L) \cap \text{supp}(q_R)|}{|\text{supp}(q_L) \cup \text{supp}(q_R)|} = \frac{|E_{1,1}|}{|E_{1,1}| + |E_{1,0}| + |E_{0,1}|}
\]
Definition: Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.

Constraints: Support, accuracy, length of the query, \(p\)-value, ...
Definition: Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.

Special cases:
- Only conjunctive queries
- Bi-directional association rules
Definition: *Given a collection of entities characterized by two sets of attributes, a redescription is a pair of queries \((q_L, q_R)\) describing the same set of entities.*
Exploration Strategies
How do we find redescriptions?

Exploration Strategies

- Exhaustive
- Heuristic
  - Mine and pair/split
  - Alternating scheme
  - Atomic updates
Exploration Strategies

*Turning CARTwheels: An Alternating Algorithm for Mining Redescriptions.*

N. Ramakrishnan, D. Kumar, B. Mishra, M. Potts, and R. F. Helm.

*Redescription Mining: Algorithms and Applications in Bioinformatics.*

D. Kumar.
Mining redescriptions with decision trees

- When one query is fixed, redescription mining turns into a **classification** problem
  - Reuse classification methods to build redescriptions
  - Require understandable queries:
    - classification and regression trees (CART)

Classification with CART’s while alternating sides
How do we find redescriptions?

Classification and Regression Trees (CART’s)

Decision tree: structure representing a succession of tests and possible classification or regression outcomes

Leaf node decision (class/value)

Other node test on an attribute’s value
How do we find redescriptions?

Classification and Regression Trees (CART’s)

Building trees iteratively

1. For each attribute calculate the entropy of split test
Classification and Regression Trees (CART’s)

Building trees iteratively

1. For each attribute calculate the entropy of split test
2. Select test with lowest entropy
3. Append new decision node for this test
4. Split the data according to the test outcome
Classification and Regression Trees (CART’s)

Building trees iteratively

1. For each attribute calculate the entropy of split test
2. Select test with lowest entropy
3. Append new decision node for this test
4. Split the data according to the test outcome
5. Recurse on each subset separately
How do we find redescriptions?

**Classification and Regression Trees (CART’s)**

Building trees iteratively

Stop if

- All entities in the subset belong to same class
- No test yields any improvement
- Maximum depth is reached
How do we find redescriptions?

Classification and Regression Trees (CART’s)

Building trees iteratively

Stop if
- All entities in the subset belong to same class
- No test yields any improvement
- Maximum depth is reached

Add leave nodes with majority class as decision
How do we find redescriptions?

The CARTwheels algorithm

Esther Galbrun
How do we find redescriptions?

The CARTwheels algorithm

- Start with some initial target
How do we find redescriptions?

The CARTwheels algorithm

- Build a classification tree over left-hand side variables
How do we find redescriptions?

The CARTwheels algorithm

- Build a classification tree over left-hand side variables

![Diagram of a classification tree with nodes Va, Vb, and Vc, and branches leading to leaves with data points.](image-url)
How do we find redescriptions?

The **CARTwheels** algorithm

- Build a classification tree over left-hand side variables

![Diagram of a classification tree with nodes Va, Vb, Vc, Vd, Ve, Vf, Vg, and Vh, with decision nodes and terminal nodes representing data points belonging to different classes.](image)
How do we find redescriptions?

**The CARTwheels algorithm**

- Relabel the entities in the leaves of the tree
How do we find redescriptions?

The CARTwheels algorithm

- Relabel the entities in the leaves of the tree

![Diagram showing a tree with nodes Va, Vb, Vc, Vd, Ve, Vf, Vg, and an arrow indicating the process of re-labeling.](image-url)
How do we find redescriptions?

The CARTwheels algorithm

- Use these new labels as the target
The CARTwheels algorithm

- Build a classification tree over right-hand side variables
The CARTwheels algorithm

- Build a classification tree over right-hand side variables
How do we find redescriptions?

The CARTwheels algorithm

- Build a classification tree over right-hand side variables
The CARTwheels algorithm

- Relabel the entities in the leaves of the tree
The CARTwheels algorithm

- Relabel the entities in the leaves of the tree
The CARTwheels algorithm

Continue to alternate...
The CARTwheels algorithm

- Continue to alternate...

How do we find redescriptions?
How do we find redescriptions?

The CARTwheels algorithm

- Obtain a pair of trees matched at the leaves
How do we find redescriptions?

The CARTwheels algorithm

- Select the branches leading to positive leaves
How do we find redescriptions?

The CARTwheels algorithm

- Compute the support
The CARTwheels algorithm

- Compute the support

How do we find redescriptions?
Exploration Strategies

Redescription Mining: Structure, Theory and Algorithms.
L. Parida and N. Ramakrishnan.
In AAAI, 2005.

Redescription Mining: Algorithms and Applications in Bioinformatics.
M. J. Zaki and N. Ramakrishnan.
In KDD, 2005.

Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining.
A. Gallo, P. Miettinen, and H. Mannila.
In SDM, 2008.
Exploration Strategies

Finding Subgroups Having Several Descriptions: Algorithms for Redescription Mining.

From Black and White to Full Color: Extending Redescription Mining Outside the Boolean World
Building redescriptions greedily

Start from a pair of singleton queries
(i.e. one variable on each side)

Try to extend the query on either side
by appending a literal using $\lor$ or $\land$

Stop if:
- The maximum length of query has been reached
- No improvement of accuracy is possible
- Previously seen redescription is repeated
- Some other constraint (support, $p$-value) is broken
How do we find redescriptions?

Building redescriptions greedily

[Diagram of redescriptions]
Building redescriptions greedily

Initial pair

How do we find redescriptions?
Building redescriptions greedily

Try to extend one side

How do we find redescriptions?
Building redescriptions greedily

Try to append $\land B$ to left hand side
How do we find redescriptions?

Building redescriptions greedily

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

Esther Galbrun
Building redescriptions greedily

...after few iterations

A B C D E

A B C D

\( (A \land D) \lor B \)

\[
J(q_L, q_R) = \frac{4}{1 + 4 + 1} = 0.66
\]
Exploration Strategies

Exhaustive

Heuristic

Mine and pair/split

Alternating scheme

Atomic updates

How do we find redescriptions?
How do we find redescriptions?

Interactive Visualisation

Interactive mining with the *Siren* interface.

http://siren.gforge.inria.fr/
Interactive Visualisation

Interactive mining with the *Siren* interface.

http://siren.gforge.inria.fr/
Interactive Visualisation

Interactive mining with the *Siren* interface.

http://siren.gforge.inria.fr/
How do we find redescriptions?

Interactive Visualisation

Interactive mining with the *Siren* interface.

http://siren.gforge.inria.fr/
Interactive Visualisation

Interactive mining with the *Siren* interface.

http://siren.gforge.inria.fr/
Redescriptions from various domains

- Computer science bibliography

Researchers with multiple publications in SoCG and CCCG conferences often collaborate with Profs M. Overmars or E. D. Demaine.
Candidates to the 2011 Finnish parliamentary election below age sixty attach little importance to the question of pension indices.
Redescriptions from various domains

- Bioclimatic niches
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)
Bioclimatic Niche Finding

European Elk

\[ ([−9.80 \leq t_{\text{Feb}}^{\text{max}} \leq 0.40] \land [12.20 \leq t_{\text{Jul}}^{\text{max}} \leq 24.60] \land [56.852 \leq p_{\text{Aug}}^{\text{avg}} \leq 136.46]) \lor [183.27 \leq p_{\text{Sep}}^{\text{avg}} \leq 238.78] \]

\[ J = 0.803 \quad |E_{1,1}| = 571 \]
Bioclimatic Niche Finding

Wood Mouse \land \text{Natterer’s Bat} \land \text{Eurasian Pygmy Shrew}

\[([3.20 \leq t_{\text{Mar}}^{\text{max}} \leq 14.50] \land [17.30 \leq t_{\text{Aug}}^{\text{max}} \leq 25.20] \land
[14.90 \leq t_{\text{Sep}}^{\text{max}} \leq 22.80] ) \lor [19.60 \leq t_{\text{Jul}}^{\text{avg}} \leq 19.956]\]

\[J = 0.621 \quad |E_{1,1}| = 677\]
Redescription Mining is a versatile and powerful data-mining tool, applicable in various domains.
Redescription Mining is a versatile and powerful data-mining tool, applicable in various domains.

Contact: esther.galbrun@inria.fr
Homepage: https://members.loria.fr/EGalbrun/web/
Siren’s website: http://siren.gforge.inria.fr/main/