Less is More: Non-Redundant Subspace Clustering

Ira Assent [○] **Emmanuel Müller** [●] Stephan Günnemann [●]
Ralph Krieger [●] Thomas Seidl [●]

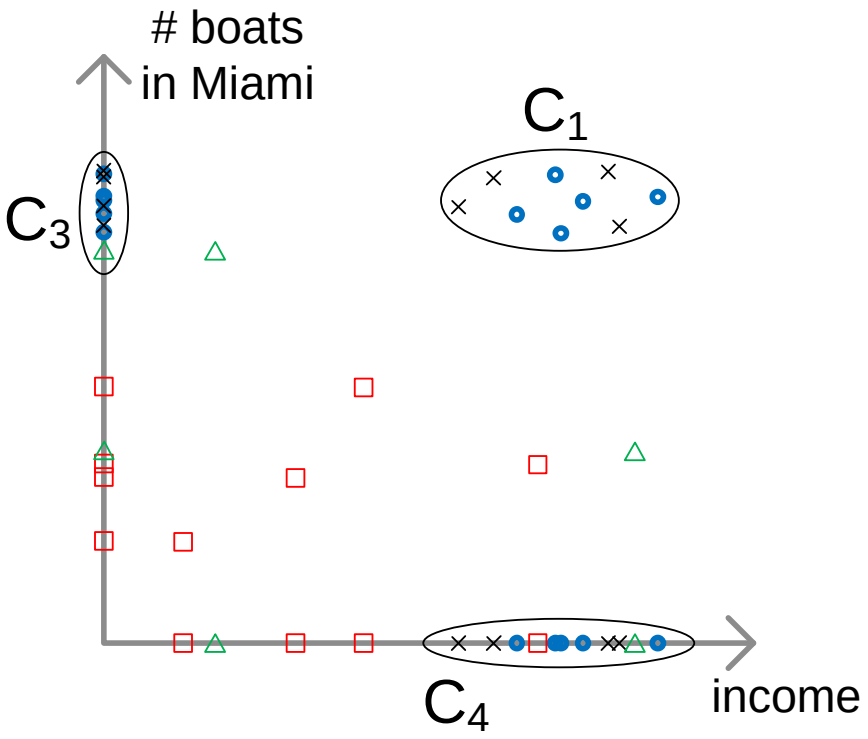
[○] Aalborg University, Denmark [●] RWTH Aachen University, Germany

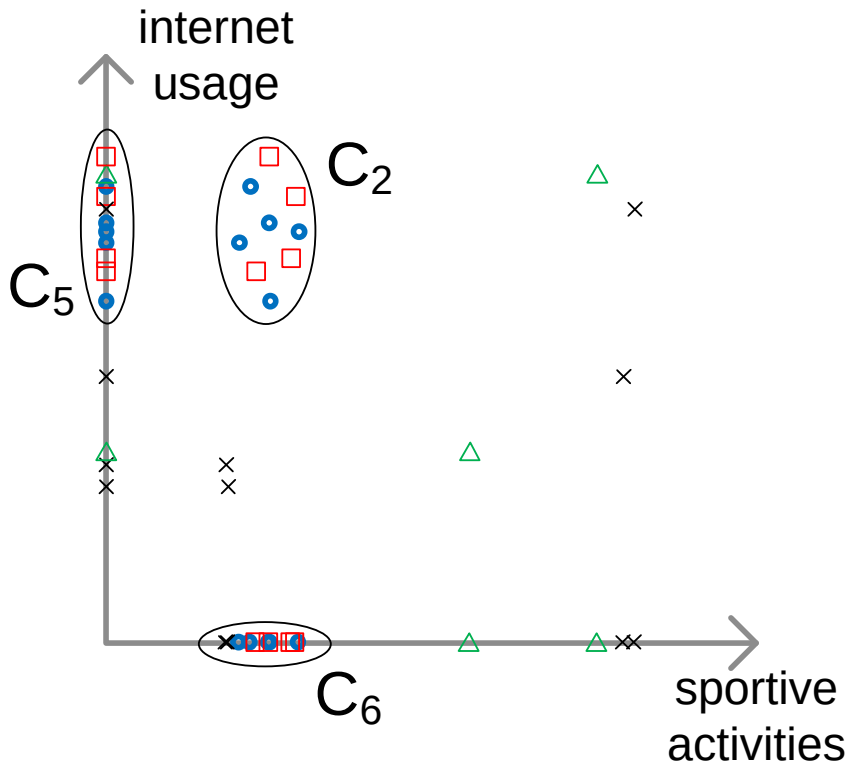
RWTH AACHEN UNIVERSITY

MultiClust Workshop at SIGKDD 2010

July 25, 2010

Detection of Non-Redundant Subspace Clusters I

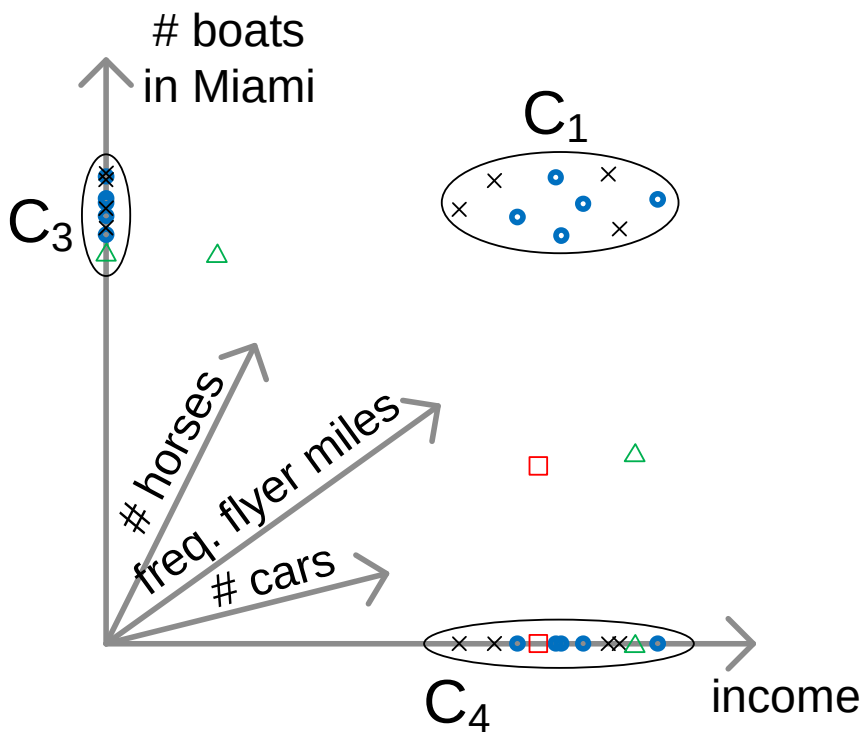




- Hidden clusters are described by different attribute sets
- Each object might be grouped in **multiple clusters**

$\Rightarrow$ Novel challenges for subspace clustering

Detection of Non-Redundant Subspace Clusters II



Subspace Cluster:

(rich; boat owner; car fan; globetrotter; horse fan)

Exp. many projections
- → (rich)
- → (boat owner)
- → (rich; globetrotter)
- ...

- Huge amount of redundant clusters

⇒ Typically number of clusters $\gg$ number of objects

⇒ **Detection of all and only non-redundant subspace clusters**

Overview

Main question

How can you use/extend non-redundant clustering ...

In this talk, we present

- A survey of our contributions so far
- The generality of our techniques
- Our open source initiatives for the community

Research questions arise in the areas of:

1. **Effective Models**
2. **Efficient Computation**
3. **Evaluation and Exploration of Results**

Notions and Related Work

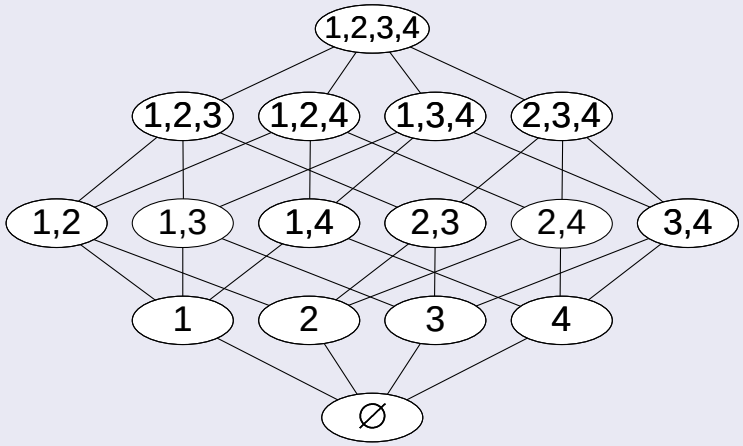
Abstract subspace clustering definition

- Definition of object set O clustered in subspace S

  $C = (O, S)$ with $O \subseteq DB, S \subseteq DIM$

- Selection of result set M a subset of all valid subspace clusters ALL

  $M = \{(O_1, S_1) \ldots (O_n, S_n)\} \subseteq ALL$



Related work

- Subspace clustering: focus on definition of (O, S)

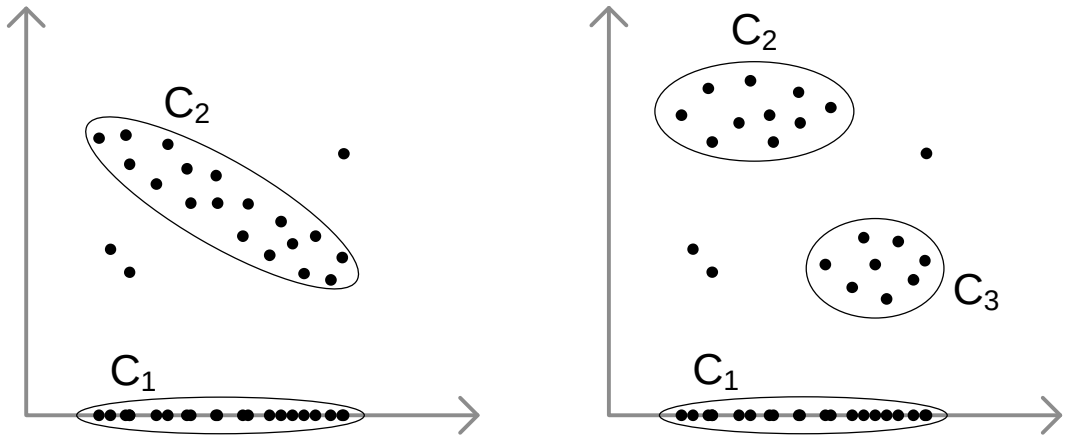
$\Rightarrow$ Output all valid subspace cluster $M = ALL$ ($\Rightarrow$ too many)

- Projected clustering: focus on definition of disjoint clusters in M

$\Rightarrow$ Unable to detect objects in multiple clusters ($\Rightarrow$ too few)

Non-Redundant Subspace Clustering Models

Select $M \subseteq ALL$: Exclude redundant subspace clusters...



Local (pairwise) redundancy elimination[1][2]

- (O, S) is non-redundant iff

$$\neg\exists(O', S') \text{ with } O' \subseteq O \wedge S' \supset S \wedge |O'| \geq R \cdot |O|$$

$\Rightarrow$ Excludes large number of redundant subspace clusters

[1] Assent, Krieger, Müller and Seidl: **DUSC: Dimensionality Unbiased Subspace Clustering**, in ICDM 2007.
[2] Assent, Krieger, Müller and Seidl: **INSCY: Indexing Subspace Clusters with In-Process-Removal of Redundancy**, in ICDM 2008.

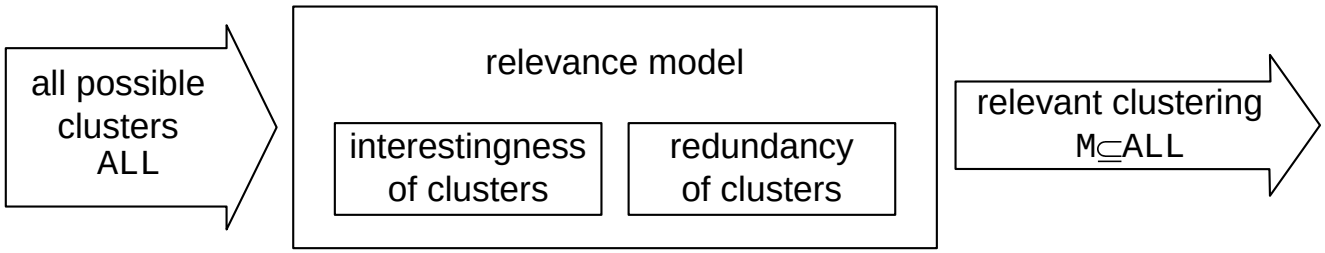
Generalization of Redundancy Elimination

Relevant subspace clustering model[3]

- Include the most interesting subspace clusters
- Exclude redundant subspace clusters

$\Rightarrow$ Provide most relevant subspace clusters in result set

$\Rightarrow$ Extract novel knowledge with each cluster



Given any definition of subspace clusters $C = (O, S)$

$\Rightarrow$ Choose **optimal** subset $M = \{C_1, \ldots, C_n\} \subseteq ALL$

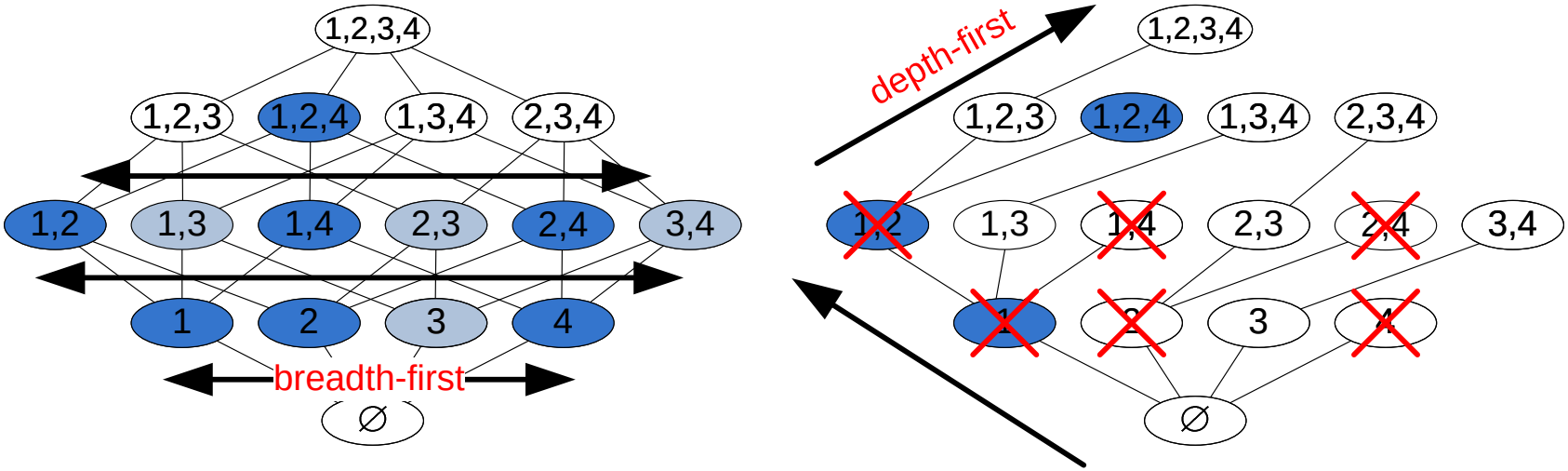
- Proof: Such an optimization is an **NP-hard problem**

[3] Müller, Assent, Günnemann, Krieger and Seidl:
Relevant Subspace Clustering: Mining the Most Interesting Non-Redundant Concepts in High Dimensional Data, in ICDM 2009.

Redundancy Pruning by Depth-First Processing

Pruning

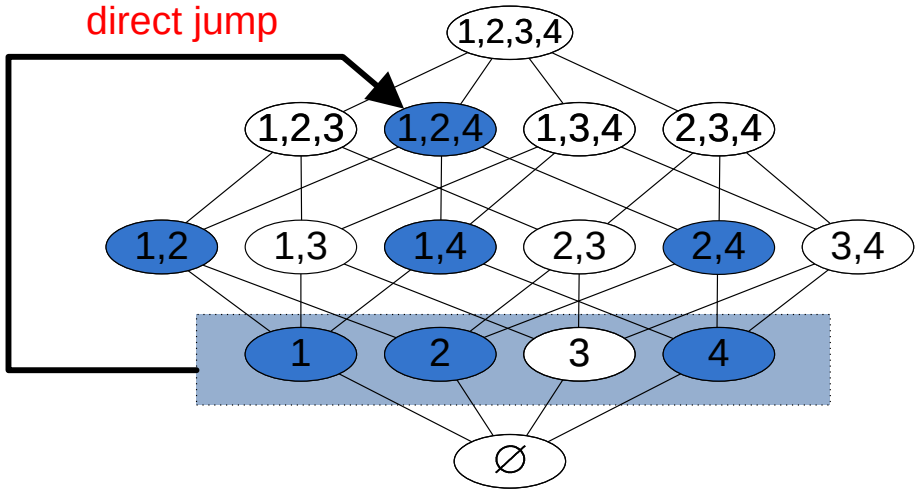
- Applicable for local redundancy (simple pairwise model)
- Enables in-process pruning of redundant clusters.

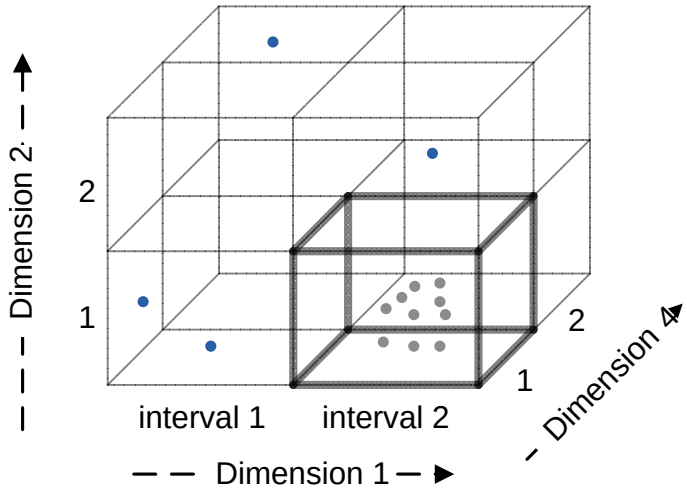


- Step-by-step processing (k-D $\rightarrow$ $(k+1)$-D subspace!)

$\Rightarrow$ Scalability to high dimensional data?

Scalable Subspace Processing





Key idea: density estimation + steered jumps

- Subspace clusters are represented by many low dimensional projections
- Use 2-D projections to estimate density in higher subspace regions[4]
- Use k-D projections to jump directly to $(k + x)$-D subspaces $[x \gg 1]$
- Best-first search: **Intelligent** steering to promising subspace regions

[4] Müller, Assent, Krieger, Günnemann and Seidl: **DensEst: Density Estimation for Data Mining in High Dimensional Spaces**, in SDM 2009.

Challenges in Evaluation and Exploration

General challenge for clustering

- No ground truth available for clustering

$\Rightarrow$ Subjective evaluation by exploration
requires **visualization** techniques and **interactive exploration** tools

$\Rightarrow$ Objective evaluations are incomparable
using different **implementations**, **databases** and **quality measures**

$\Rightarrow$ We provide broad evaluation study & interactive exploration framework

Evaluation Study[5]

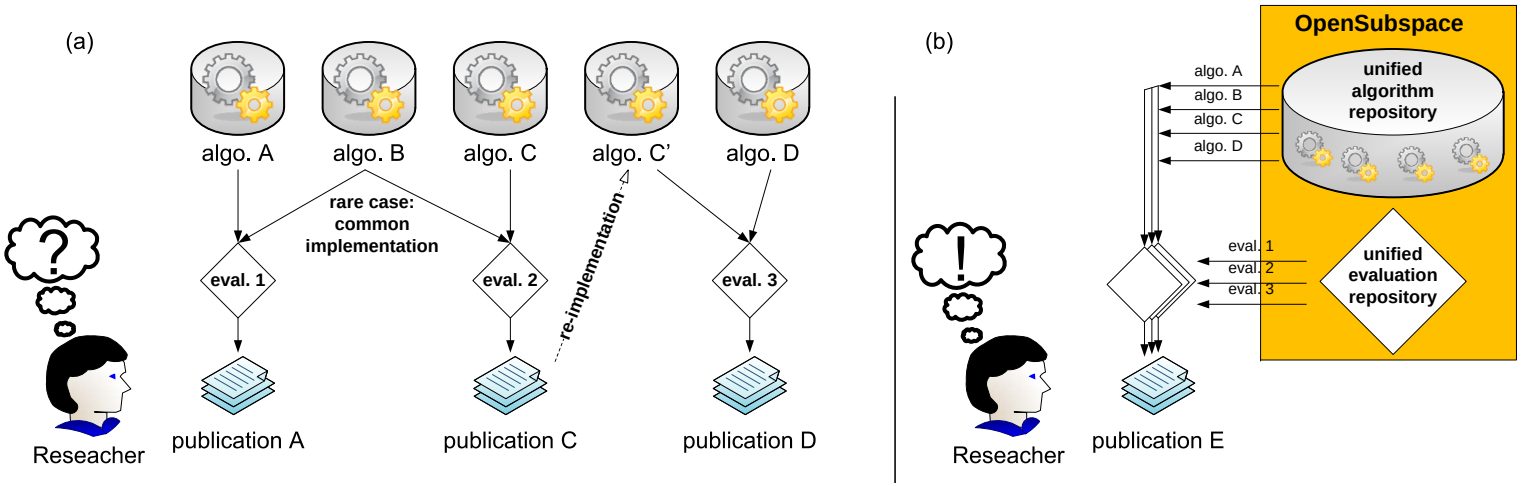
- Characterization of major paradigms
- Providing comparable baseline implementations
- Evaluation based on broad set of
data sets, **quality measures** and **parameter settings**

[5] Müller, Günnemann, Assent and Seidl: **Evaluating Clustering in Subspace Projections of High Dimensional Data**, in VLDB 2009.

Open Source Framework

OpenSubspace framework

- Framework for research, education and application [6][7][8][9]
- Baselines for algorithm and evaluation measure development



http://dme.rwth-aachen.de/OpenSubspace/

[6] Müller, Assent, Krieger, Jansen and Seidl: **Morpheus: Interactive Exploration of Subspace Clustering**, in KDD 2008.
[7] Assent, Müller, Krieger, Jansen and Seidl: **Pleiades: Subspace Clustering and Evaluation**, in PKDD 2008.
[8] Günnemann, Färber, Kremer, Seidl: **CoDA: Interactive Cluster Based Concept Discovery**, in VLDB 2010
[9] Müller, Schiffer, Gerwert, Hannen, Jansen, Seidl: **SOREX: Subspace Outlier Ranking Exploration Toolkit**, in PKDD 2010.

Conclusion and Future Work

Subspace clustering is still an emerging research field...

Is the basis for a lot of further research

- **Alternative** subspace clustering
- Evaluation **measures** for subspace clustering
- **Benchmark databases** for subspace clustering
- . . .

Conclusion and Future Work

Subspace clustering is still an emerging research field...

Is the basis for a lot of further research

- **Alternative** subspace clustering
- Evaluation **measures** for subspace clustering
- **Benchmark databases** for subspace clustering
- . . .

Thank you for your attention.

Questions?