Evaluating Clustering in Subspace Projections of High Dimensional Data

Emmanuel Müller • Stephan Günnemann • **Ira Assent** ◦ Thomas Seidl •

• RWTH Aachen University, Germany

◦ Aalborg University, Denmark

VLDB 2009
Lyon, France

Overview

Clustering

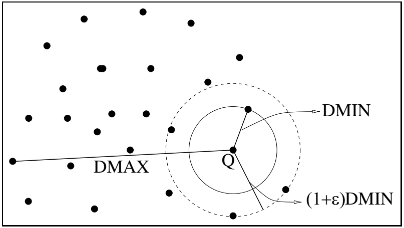
- Group **similar** objects, separate dissimilar ones
- Usually similarity given by means of a distance function

Problems in high dimensional data

- "Curse of Dimensionality"[1]

⇒ Distances grow more and more alike

⇒ No meaningful clusters

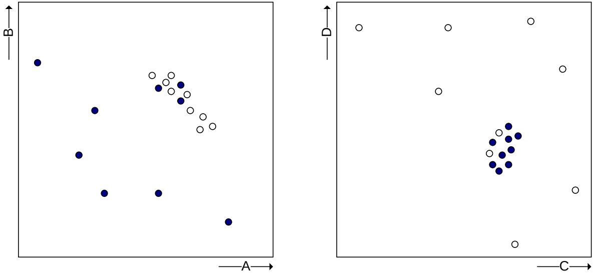


[1]

[1] Beyer, Goldstein, Ramakrishnan and Shaft, **When is nearest neighbors meaningful**, in ICDT 1999.

Clustering High Dimensional Data

- Clusters appear in subspaces of the data
- Global dimensionality reduction techniques find a **single** projection



General challenge

- Each cluster has its own relevant dimensions

Clustering in Subspace Projections

Subspace cluster

A cluster C in a subspace projection S is

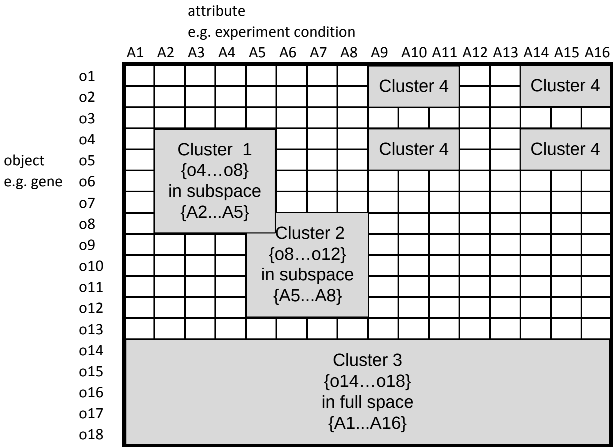
$$C = (O, S) \text{ with } O \subseteq DB, S \subseteq D$$

Subspace clustering

A clustering result R of k clusters is a set of clusters

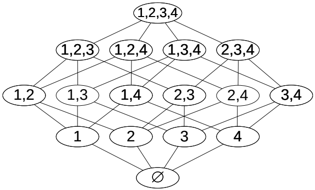
$$R = \{C_1, \ldots, C_k\}, C_i = (O_i, S_i) \text{ for } i = 1 \ldots k$$

Example


attribute
e.g. experiment condition
A1 A2 A3 A4 A5 A6 A7 A8 A9 A10 A11 A12 A13 A14 A15 A16
object
e.g. gene
o1
o2
o3
o4
o5
o6
o7
o8
o9
o10
o11
o12
o13
o14
o15
o16
o17
o18
Cluster 4
Cluster 4
Cluster 4
Cluster 4
Cluster 1
{o4…o8}
in subspace
{A2...A5}
Cluster 2
{o8…o12}
in subspace
{A5...A8}
Cluster 3
{o14…o18}
in full space
{A1...A16}

Challenges in Cluster Detection

- Clusters in arbitrary subsets of dimensions
- Exponential number of possible subspaces



- **Various approaches** have been proposed
- Comparison has been done only for a small subset of approaches

Our goal

⇒ Systematic evaluation and comparison

Paradigms

Algorithmic view often found in the literature:

- Projected (partitioning) vs. subspace (overlapping) clustering

Here: new model-centered view

Cell-based subspace clustering

- Discretization; efficient detection of dense grid cells

Density-based subspace clustering

- Dense areas separated by sparse areas

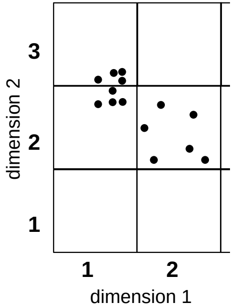
Clustering oriented

- Optimize the overall clustering result

Note: we do not study application specific approaches (e.g. bioinformatics)

Cell-Based Approaches

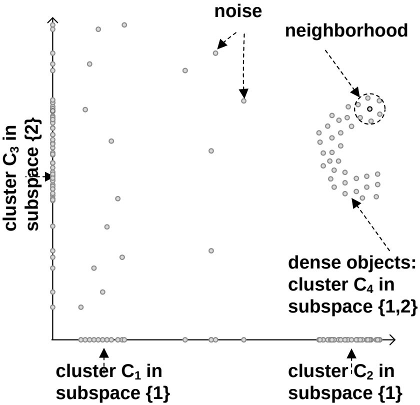
- Discretization via fixed or variable grid structure
  $\rightarrow$ Approximation
- Cluster definition based on object count per cell
  $\rightarrow$ Cluster result is a set of cells with $(O \geq \tau)$



- Studied approaches:
  CLIQUE [SIGMOD 1998], DOC [SIGMOD 2002], MINECLUS [ICDM 2003], SCHISM [ICDM 2004]

Density-Based Approaches

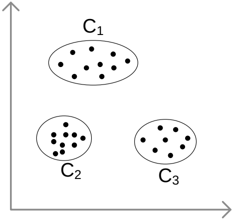
- Based on DBSCAN [KDD 1996]
- Dense clusters separated by sparse regions
- Cluster definition based on density in neighborhood of each object



- Studied approaches:
  SUBCLU [SDM 2004], FIRES [ICDM 2005], INSCY [ICDM 2008]

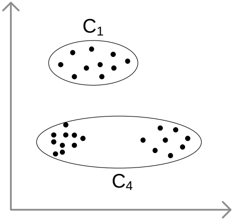
Clustering Oriented Approaches

- Focus on entire clustering result
  → Objective function
- Global optimization
  → Control resulting cluster set
  → No direct influence on each individual cluster



numClusters(C_1,C_2,C_3) = 3

- Studied approaches:
  PROCLUS [SIGMOD 1999],
  P3C [ICDM 2006],
  STATPC [KDD 2008]



numClusters(C_1,C_4) = 2

Evaluation of Clustering

Challenges

- Lack of Ground Truth
  - Clustering searches for yet **unknown patterns**
- Cluster Analysis by Domain Expert
  - Practical usefulness of results; not objective
- Evaluation Measures
  Resort to measures as in classification analysis
  - Assumes that class labels reflect ideal clustering
  - Class might be split in two; classes might share common structures

Overview of Evaluation Measures

Evaluation paradigms

- Simple Measures (assume no additional knowledge)
  - Coverage (objects) and cluster distribution (dimensions)
- Enhanced measures (assume class labels)
  - Entropy, precision, recall, F1, classification accuracy
- Subspace measures (assume subspace ground truth)
  - Cluster Error (CE) and Relative Non-Intersecting Area (RNIA)

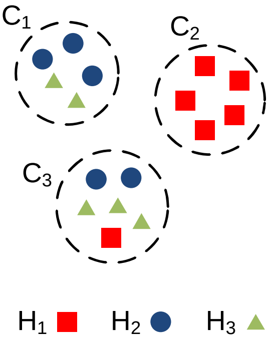
Enhanced Measures I

F1 value

- Harmonic mean of recall and precision
  - All objects of hidden cluster detected?
  - How accurately detected?
- Map each found cluster C_i to best covered hidden cluster H_j

$$\frac{|O_i \cap O_H|}{|O_H|} \geq \frac{|O_i \cap O_{H_j}|}{|O_{H_j}|} \quad \forall j \in \{1, \ldots, m\}$$

$\Rightarrow$ High F1 value: detected object grouping corresponds to hidden groups



Enhanced Measures II

Entropy

- Homogeneity / purity of found clusters

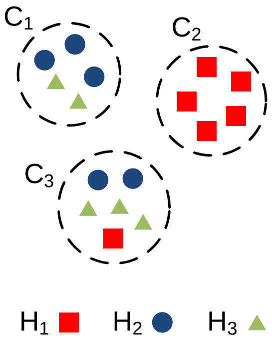
$$E(C) = -\sum_{i=1}^{m} p(H_i|C) \cdot \log(p(H_i|C))$$

- Weighted average over all entropy values

Accuracy

$$\frac{|\text{correctly predicted objects}|}{|\text{all objects}|}$$

- Build classifier on found clusters
- High accuracy: clusters generalize data well



Subspace Measures

*Are **subobjects** (in correct projections) detected?*

RNIA measure

- I intersection of all hidden and found clusters
- U union of all hidden and found clusters

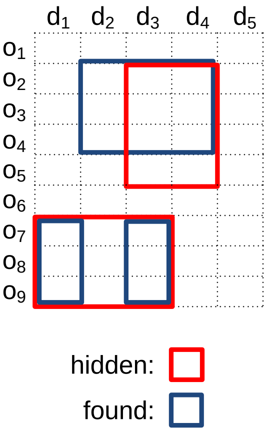
$$RNIA = (U - I)/U$$

- Split-up of clusters not considered

CE measure

- 1-to-1 mapping of hidden and found clusters
- $\bar{I}$ maximum intersection of mapped pairs

$$CE = (U - \bar{I})/U$$



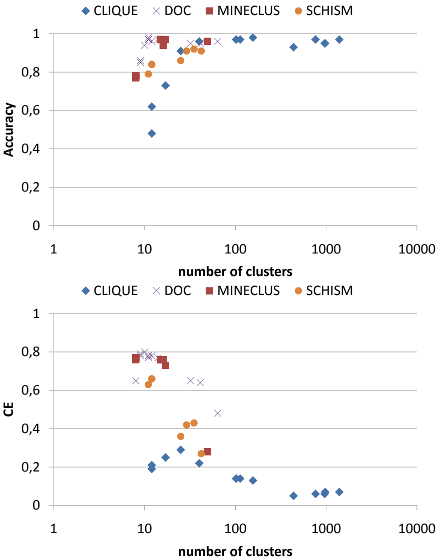
Experiment Setup

Data sets

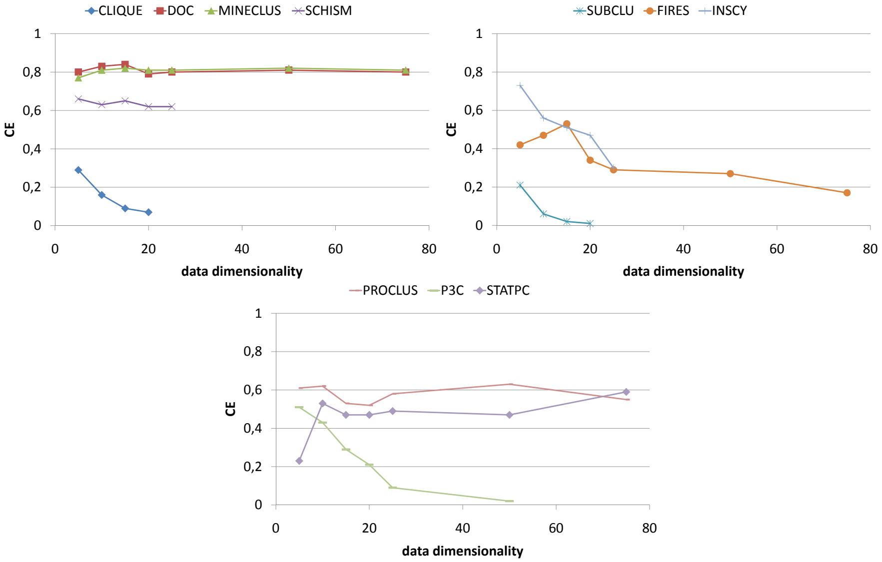
- Generated **synthetic data** with 10 hidden subspace clusters with a dimensionality of 50%, 60% and 80% of different dimensionalities
- Benchmark **real world data** from UCI ML repository

Fair comparison

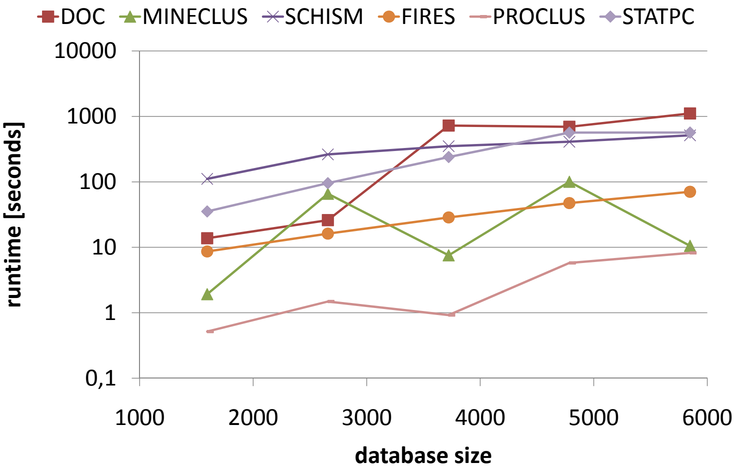
- **Parametrization:** broad range of parameter settings
  $\rightarrow$ try to find for each algorithm best parameters on each data set
- **Broad evaluation:** enormous amount of experiment runs
  (23 data sets $\times$ 10 algorithms $\times$ on average 100 parameter settings)
  $\rightarrow$ restricted runtime for each run to 30 minutes
- **Evaluation framework:**
  http://dme.rwth-aachen.de/OpenSubspace/



Different measures for cell based approaches



Scalability: **CE measure** vs. **database dimensionality**



Scalability: **Runtime** vs. **database size**

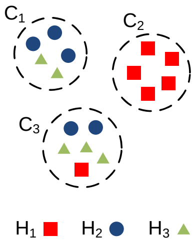
Glass (size: 214; dim: 9)	F1		Accuracy		CE		RNIA		Entropy		Coverage		NumClusters		AvgDim		Runtime	
	max	min	max	min	max	min	max	min	max	min	max	min	max	min	max	min	max	min
CLIQUE	0,51	0,31	0,67	0,50	0,02	0,00	0,06	0,00	0,39	0,24	1,00	1,00	6169	175	5,4	3,1	411195	1375
DOC	0,74	0,50	0,63	0,50	0,23	0,13	0,93	0,33	0,72	0,50	0,93	0,91	64	11	9,0	3,3	23172	78
MINECLUS	0,76	0,40	0,52	0,50	0,24	0,19	0,78	0,45	0,72	0,46	1,00	0,87	64	6	7,0	4,3	907	15
SCHISM	0,46	0,39	0,63	0,47	0,11	0,04	0,33	0,20	0,44	0,38	1,00	0,79	158	30	3,9	2,1	313	31
SUBCLU	0,50	0,45	0,65	0,46	0,00	0,00	0,01	0,01	0,42	0,39	1,00	1,00	1648	831	4,9	4,3	14410	4250
FIRES	0,30	0,30	0,49	0,49	0,21	0,21	0,45	0,45	0,40	0,40	0,86	0,86	7	7	2,7	2,7	78	78
INSCY	0,57	0,41	0,65	0,47	0,23	0,09	0,54	0,26	0,67	0,47	0,86	0,79	72	30	5,9	2,7	4703	578
PROCLUS	0,60	0,56	0,60	0,57	0,13	0,05	0,51	0,17	0,76	0,68	0,79	0,57	29	26	8,0	2,0	375	250
P3C	0,28	0,23	0,47	0,39	0,14	0,13	0,30	0,27	0,43	0,38	0,89	0,81	3	2	3,0	3,0	32	31
STATPC	0,75	0,40	0,49	0,36	0,19	0,05	0,67	0,37	0,88	0,36	0,93	0,80	106	27	9,0	9,0	1265	390

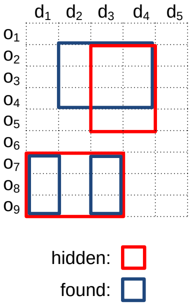
Real world example: Glass data set

Discussion

Measures

- Applicability governed by available ground truth
- Accuracy ignores split-up of clusters; entropy biased towards high dimensionality → F1 more meaningful
- RNIA and CE consider relevant dimensions, penalize large result sets → CE measure considers split-up of clusters

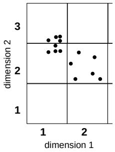


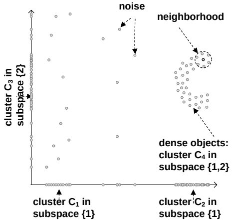


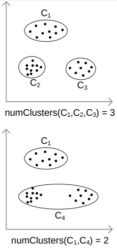
Discussion II

Clustering paradigms

- Density-based approaches do not scale to very high dimensional data
- Clustering oriented approaches are affected by noisy data
- Recent cell-based MINECLUS efficient and effective
- Basic clustering oriented PROCLUS performs well
- Basic approaches CLIQUE (cell-based) and SUBCLU (density-based) produce tremendously large result set → recent approaches of these paradigms enhanced quality and efficiency; top results only in few cases







Conclusion

- Experimental evaluation of subspace clustering and evaluation measures
  - Important comparative study for subspace clustering research
  - Characterization of measures and different paradigms
  - Helpful for further research
- Results: good overall performance of
  - Cell-based MINECLUS
  - Clustering-oriented PROCLUS
  - Enhanced measure F1
  - Subspace measure CE
- For comparison, repeatability, or further research
  http://dme.rwth-aachen.de/OpenSubspace/evaluation
  → full information on all parametrizations, results, data sets, and download of open source implementation in WEKA

Conclusion

- Experimental evaluation of subspace clustering and evaluation measures
  - Important comparative study for subspace clustering research
  - Characterization of measures and different paradigms
  - Helpful for further research
- Results: good overall performance of
  - Cell-based MINECLUS
  - Clustering-oriented PROCLUS
  - Enhanced measure F1
  - Subspace measure CE
- For comparison, repeatability, or further research
  http://dme.rwth-aachen.de/OpenSubspace/evaluation
  → full information on all parametrizations, results, data sets, and download of open source implementation in WEKA

Thank you for your attention.

Questions?