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

1. Introduction
2. Clustering Paradigms
3. Evaluation Setup
4. Experiments
5. Conclusion
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

---

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

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

<table>
<thead>
<tr>
<th>Attribute</th>
<th>e.g. experiment condition</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>A2</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Object</th>
<th>e.g. gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>o1</td>
<td>o2</td>
</tr>
</tbody>
</table>

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\}

Cluster 4: \{o4…o8\} in subspace \{A2…A5\}

Cluster 4: \{o8…o12\} in subspace \{A5…A8\}

Cluster 4: \{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 → Approximation
- Cluster definition based on object count per cell → 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

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

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

⇒ 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
- |correctly predicted objects| / |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

\[
RNIA = \frac{(U - I)}{U}
\]

- Split-up of clusters not considered

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

\[
CE = \frac{(U - \bar{I})}{U}
\]
# 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
  → try to find for each algorithm best parameters on each data set
- **Broad evaluation**: enormous amount of experiment runs
  (23 data sets × 10 algorithms × on average 100 parameter settings)
  → restricted runtime for each run to 30 minutes
Different measures for cell based approaches
Scalability: **CE measure vs. database dimensionality**
Scalability: **Runtime vs. database size**
Glass (size: 214; dim: 9)

<table>
<thead>
<tr>
<th>Method</th>
<th>F1</th>
<th>Accuracy</th>
<th>CE</th>
<th>RNIA</th>
<th>Entropy</th>
<th>Coverage</th>
<th>NumClusters</th>
<th>AvgDim</th>
<th>Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>max</td>
<td>min</td>
<td>max</td>
<td>min</td>
<td>max</td>
<td>min</td>
<td>max</td>
<td>min</td>
<td>max</td>
</tr>
<tr>
<td>CLIQUE</td>
<td>0.51</td>
<td>0.31</td>
<td>0.67</td>
<td>0.50</td>
<td>0.02</td>
<td>0.00</td>
<td>1.00</td>
<td>1.00</td>
<td>411195</td>
</tr>
<tr>
<td>DOC</td>
<td>0.74</td>
<td>0.50</td>
<td>0.63</td>
<td>0.50</td>
<td>0.23</td>
<td>0.13</td>
<td>0.93</td>
<td>0.33</td>
<td>64</td>
</tr>
<tr>
<td>MINECLUS</td>
<td>0.76</td>
<td>0.40</td>
<td>0.52</td>
<td>0.50</td>
<td>0.24</td>
<td>0.19</td>
<td>0.78</td>
<td>0.45</td>
<td>158</td>
</tr>
<tr>
<td>SCHISM</td>
<td>0.46</td>
<td>0.39</td>
<td>0.63</td>
<td>0.47</td>
<td>0.11</td>
<td>0.04</td>
<td>0.33</td>
<td>0.20</td>
<td>1.00</td>
</tr>
<tr>
<td>SUBCLU</td>
<td>0.50</td>
<td>0.45</td>
<td>0.65</td>
<td>0.46</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.01</td>
<td>64</td>
</tr>
<tr>
<td>FIRES</td>
<td>0.30</td>
<td>0.30</td>
<td>0.49</td>
<td>0.49</td>
<td>0.21</td>
<td>0.21</td>
<td>0.45</td>
<td>0.45</td>
<td>1648</td>
</tr>
<tr>
<td>INSCY</td>
<td>0.57</td>
<td>0.41</td>
<td>0.65</td>
<td>0.47</td>
<td>0.23</td>
<td>0.09</td>
<td>0.54</td>
<td>0.26</td>
<td>7</td>
</tr>
<tr>
<td>PROCLUS</td>
<td>0.60</td>
<td>0.56</td>
<td>0.60</td>
<td>0.57</td>
<td>0.13</td>
<td>0.05</td>
<td>0.51</td>
<td>0.17</td>
<td>72</td>
</tr>
<tr>
<td>P3C</td>
<td>0.28</td>
<td>0.23</td>
<td>0.47</td>
<td>0.39</td>
<td>0.14</td>
<td>0.13</td>
<td>0.30</td>
<td>0.27</td>
<td>29</td>
</tr>
<tr>
<td>STATPC</td>
<td>0.75</td>
<td>0.40</td>
<td>0.49</td>
<td>0.36</td>
<td>0.19</td>
<td>0.05</td>
<td>0.67</td>
<td>0.37</td>
<td>3</td>
</tr>
</tbody>
</table>

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](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
  
  [Link](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?