The Blind Men and the Elephant

Arthur Zimek

Ludwig-Maximilians-Universität München

Talk at NII, Tokyo
23.03.2015
The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms
Clustering Uncertain Data
Conclusions
References
The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms

Clustering Uncertain Data

Conclusions

References
Clustering Paradigms

- Subspace Clustering:
  Find clusters in subspaces of the data. The subspaces where clusters reside are previously unknown.

- Ensemble Clustering:
  Derive various clustering results, unify different results to a single, supposedly more reliable result.

- Alternative Clustering:
  Given some clustering result, find a different clustering (can also be seen as a way of constraint clustering).

- Multi-View Clustering:
  Discover different clusterings in different subspaces.
Outline

Clustering Paradigms
  Subspace Clustering
  Ensemble Clustering
  Alternative Clustering
  Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions
Relevant and Irrelevant Attributes

- **feature relevance**
  - subset of features relevant for clustering
  - cluster can be identified in this subspace only

- **feature correlation**
  - a subset of features can be correlated
  - relevant subspace is arbitrarily oriented
Local Feature Relevance

different feature subsets/correlations are relevant for different clusters/outliers

3D data set with four clusters

projection on x/z (relevant for red/green cluster)

projection on x/y

projection on y/z (relevant for blue/purple cluster)
Problem Setting

Clustering in High-Dimensional Data

Search for clusters in (in general arbitrarily oriented) subspaces of the original feature space

Challenges:

▸ Find the correct subspace of each cluster
  ▸ Search space:
    ▶ all possible arbitrarily oriented subspaces of a feature space
    ▶ infinite

▸ Find the correct cluster in each relevant subspace
  ▸ Search space:
    ▶ “Best” partitioning of points
    ▶ NP-complete
circular dependency:
► in order to determine the correct subspace of a cluster, we need to know (at least) some cluster members
► in order to determine the correct cluster membership, we need to know the subspaces of all clusters

solution strategy
► integrate subspace search into clustering process
► requires heuristics to solve
  ► the clustering problem and
  ► the subspace search problem simultaneously
Assumptions and Categories of Approaches

- one common assumption to restrict the search space: we look for clusters in axis-parallel subspaces only
- subspace search traversal can
  - start from one-dimensional spaces and combine them (bottom-up)
  - start in the full-dimensional space and deselect attributes (top-down)
- result set of clusters can
  - partition the data into disjoint clusters (“projected clustering”)
  - find all clusters in all subspaces (possibly with a huge overlap/redundancy) (“subspace clustering”)
- both taxonomies are not strict
- other assumptions lead to other approaches (correlation clustering, pattern-based/ co-/ bi-clustering [Kriegel et al., 2009, Madeira and Oliveira, 2004])
Summary

- Rich literature – some surveys by Kriegel et al. [2009, 2012], Sim et al. [2013], Zimek [2013], Zimek et al. [2014].
- Essential point here: many approaches are interested in finding potentially different clustering results in different subspaces.
- Problem: very similar clusters might be found over and over again in different subspaces (redundancy).
Outline

Clustering Paradigms
- Subspace Clustering
- Ensemble Clustering
- Alternative Clustering
- Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions
Assume a binary classification problem (e.g., “does some item belong to class ‘A’ or to class ‘B’?”)

- in a “supervised learning” scenario (i.e., we have examples for ‘A’ and examples for ‘B’), we can learn a model (i.e., train a classifier)
- some classifier (model) decides with a certain accuracy
- error rate of the classifier: in how many cases (percentage) is the decision wrong?
- “ensemble”: ask several classifiers, combine their decision (e.g., majority vote)
Ensembles for Classification

- the ensemble will be much more accurate than its components, *if*
  - the components decide independently,
  - and each component decides more accurate than a coin.
- In supervised learning, the literature provides a well developed theory for ensembles.
The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms
Subspace Clustering
Ensemble Clustering
Alternative Clustering
Multi-View Clustering
Summary
Clustering Uncertain Data
Conclusions
References

Error-Rate of Ensembles for Classification

\[
\bar{p}(k, p) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} p^i (1 - p)^{k-i}
\]

(relates to Condorcet’s Jury theorem [Marquis de Condorcet, 1785])
Error-rate and similar concepts are not clearly applicable in clustering, yet the basic intuition is transferred from classification to clustering:

**Motivation:**

All the ensemble members are committing errors, hopefully not too many, but on different cases, *if* the members are independent, i.e., diverse.

By combining such diverse base methods into an ensemble, the different errors should level out.

Therefore the basic steps are:

- construct diverse enough (but “accurate”) clusterings for the same data set
- combine the diverse decisions into an ensemble
Error-rate and similar concepts are not clearly applicable in clustering, yet the basic intuition is transferred from classification to clustering:

Motivation:
All the ensemble members are committing errors, hopefully not too many, but on different cases, if the members are independent, i.e., diverse.
By combining such diverse base methods into an ensemble, the different errors should level out.

Therefore the basic steps are:
- construct diverse enough (but “accurate”) clusterings for the same data set
- combine the diverse decisions into an ensemble
Weak Theory in Clustering

Error-rate and similar concepts are not clearly applicable in clustering, yet the basic intuition is transferred from classification to clustering:

Motivation:
All the ensemble members are committing errors, hopefully not too many, but on different cases, *if* the members are independent, i.e., diverse. By combining such diverse base methods into an ensemble, the different errors should level out.

Therefore the basic steps are:

- construct diverse enough (but “accurate”) clusterings for the same data set
- combine the diverse decisions into an ensemble
Weak Theory in Clustering

Error-rate and similar concepts are not clearly applicable in clustering, yet the basic intuition is transferred from classification to clustering:

**Motivation:**

All the ensemble members are committing errors, **hopefully not too many**, but on different cases, *if* the members are independent, i.e., diverse.

By combining such diverse base methods into an ensemble, the different errors should level out.

Therefore the basic steps are:

- construct diverse enough (but “accurate”) clusterings for the same data set
- combine the diverse decisions into an ensemble
Combination procedures for cluster ensembles are relatively well explored [Topchy et al., 2004, 2005, Long et al., 2005, Fred and Jain, 2005, Caruana et al., 2006, Domeniconi and Al-Razgan, 2009, Gullo et al., 2009, Hahmann et al., 2009, Singh et al., 2010].

There are only a few studies on the impact of diversity in cluster ensembles, such as the work by Kuncheva and Hadjitodorov [2004], Brown et al. [2005], Hadjitodorov et al. [2006], Hadjitodorov and Kuncheva [2007].
Heuristics for Diversity

Possible heuristics to obtain diverse results (as discussed by Strehl and Ghosh [2002]):

- non-identical sets of features
- non-identical sets of objects
- different clustering algorithms

First strategy clearly related to subspace clustering

Has been pursued in many approaches, e.g., by Fern and Brodley [2003], Topchy et al. [2005], Bertoni and Valentini [2005], but:

- typically, projections are random
- not different solutions are sought in different subspaces, but true clusters are supposed to be more or less equally apparent in different randomized projections
It is probably interesting to account for the possibility of different, yet meaningful, clusters in different subspaces, so ensembles should unify only similar (?) clusters.

Possibly, subspace clustering can benefit from advanced clustering diversity measures in ensemble clustering [Strehl and Ghosh, 2002, Fern and Brodley, 2003, Hadjitodorov et al., 2006, Gionis et al., 2007, Fern and Lin, 2008], in order to avoid redundancy.
Outline

Clustering Paradigms
Subspace Clustering
Ensemble Clustering
Alternative Clustering
Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions
Constraint Clustering

- classification (supervised learning): a model is learned based on complete information about the class structure of a (training) data set
- clustering (unsupervised learning): a model is fit to a data set without using prior information
- semi-supervised clustering is using some information, e.g., some objects may be labeled
- this partial class information is used to derive must-link constraints or cannot-link constraints [Klein et al., 2002, Bade and Nürnberger, 2008, Basu et al., 2008, Davidson and Ravi, 2009, Lelis and Sander, 2009, Zheng and Li, 2011, Campello et al., 2013, Pourrajabi et al., 2014]
Alternative Clustering

Motivation by Gondek and Hofmann [2005]:
“users are often unable to positively describe what they are looking for, yet may be perfectly capable of expressing what is not of interest to them”

- constraint of non-redundancy, given some clustering
- objects clustered together in the given clustering should not be clustered together in a new clustering [Gondek and Hofmann, 2004, Bae and Bailey, 2006, Jain et al., 2008, Davidson et al., 2010, Niu et al., 2010, Dang and Bailey, 2010, Dang et al., 2012]
- a common heuristic to get a different clustering: use different subspaces [Qi and Davidson, 2009, Günneemann et al., 2009] – allowing some redundancy?
- survey: ?
Outline

Clustering Paradigms
  Subspace Clustering
  Ensemble Clustering
  Alternative Clustering

Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions
Semantically Different Subspaces

- some methods treat different subsets of attributes or different data representations separately based on a different semantic of these subsets
  - example: color features vs. texture features

- distances in the combined color-texture space are meaningless
Different Goals of Multi-View Clustering


- find *different* concepts realized in different feature subsets (the ‘subspace clustering’ idea – but subspaces are semantically meaningful) [Cui et al., 2007, Jain et al., 2008]
  - special case of alternative clustering (constraint: orthogonality of the subspaces) [Dang et al., 2012]
  - or special case of subspace clustering allowing maximal overlap yet seeking minimally redundant clusters by accommodating different concepts [Günnemann et al., 2009]
The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms
Subspace Clustering
Ensemble Clustering
Alternative Clustering
Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions

Clustering Paradigms
Subspace Clustering
Ensemble Clustering
Alternative Clustering
Multi-View Clustering

Summary

Clustering Uncertain Data

Conclusions
Different Meaningful Clusterings?

The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms
Subspace Clustering
Ensemble Clustering
Alternative Clustering
Multi-View Clustering

Summary

Clustering Uncertain Data
Conclusions
References
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: ALOI [Geusebroek et al., 2005]
Different Meaningful Clusterings?

Example: Pendigits [Bache and Lichman, 2013]

observation by Färber et al. [2010]
Clustering Paradigms

- Multi-view clustering I
  - High redundancy
  - Many results
  - Single result

- Multi-view clustering II
  - Low redundancy
  - Alternative clustering
    - Problem: balance between admissible overlap of clusters and difference between concepts
  - Many results

- Subspace clustering
  - Low redundancy
  - Ensemble clustering
Cold Water
uncertain data can occur in very different scenarios, such as

- sensor readings
- recognition and parsing
- predictions and extrapolations
- machine learning tasks
- etc. . .
Example: Geo-Spatial Data

geo-spatial data may be uncertain due to

▶ erroneous/inexact GPS readings
▶ triangulation errors
▶ human error
▶ etc. . .

uncertain spatial data may be represented

▶ continuously

▶ discretely
Dealing with Uncertainty

approach 1: clean
(i.e., remove uncertainty)

pro:
▶ can use traditional DBMS (and clustering...)

con:
▶ cleaning non-trivial
▶ can results be trusted?

approach 2: manage
(i.e., keep uncertainty)

pro:
▶ preserves information
▶ can provide confidence

con:
▶ specialized DBMS (and clustering...)

pro:
▶ can use traditional DBMS (and clustering...)

con:
▶ cleaning non-trivial
▶ can results be trusted?
Possible Worlds Semantics

Cleaning non-trivial: e.g., constraints on data may not be fulfilled when using aggregates.

- GPS readings around a lake
- the mean of all readings is not a valid position

solution: sampling possible worlds
Sampling of Possible Worlds

The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms

Clustering Uncertain Data

Conclusions

References
Sampling of Possible Worlds
Sampling of Possible Worlds
Sampling of Possible Worlds
Sampling of Possible Worlds
Solution space: metric space of ARI-distances between clustering solutions
The Blind Men and the Elephant

Arthur Zimek

Outline

Clustering Paradigms

Clustering Uncertain Data

Conclusions
Conclusions

- Multi-view clustering I
  - High redundancy
  - Many results
  - Problem: balance between admissible overlap of clusters and difference between concepts
- Multi-view clustering II
- Single result
- Low redundancy
- Ensemble clustering
- Subspace clustering
- Alternative clustering

References


http://doi.org/10.1145/2623330.2623725
The Blind Men and the Elephant

Arthur Zimek

Clustering Paradigms
Clustering Uncertain Data
Conclusions
References

Thank you for your attention!
References I


A. Bertoni and G. Valentini. Ensembles based on random projections to improve the accuracy of clustering algorithms. In *16th Italian Workshop on Neural Nets (WIRN), and International Workshop on Natural and Artificial Immune Systems (NAIS), Vietri sul Mare, Italy*, pages 31–37, 2005. doi: 10.1007/11731177_5.


References IV


References V


References VII


