Contents

Evaluation of Clustering Results

Ein paar Grundgedanken:

- Clustering: Anwendung eines Verfahrens auf ein konkretes Problem (einen bestimmten Datensatz, über den man Neues erfahren möchte).

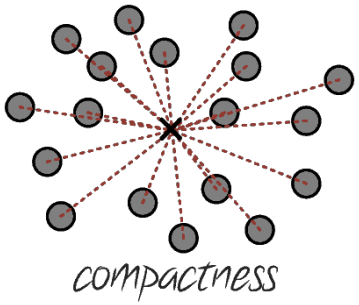
  Zur Erinnerung: KDD ist der Prozess der (semi-) automatischen Extraktion von Wissen aus Datenbanken, das gültig, bisher ***unbekannt*** und potentiell nützlich ist.

- Frage: Was kann man mit den Ergebnissen anfangen?
  - nicht unbedingt neue Erkenntnisse, aber gut überprüfbare
  - Anwendung auf Daten, die bereits gut bekannt sind
  - Anwendung auf künstliche Daten, deren Struktur by design bekannt ist
  - Frage: Werden Eigenschaften/Strukturen gefunden, die der Algorithmus nach seinem Model finden sollte? Besser als andere Algorithmen?
  - Überprüfbarkeit alleine ist aber fragwürdig!
- grundsätzlich: Clustering ist unsupervised
  - ein Clustering ist nicht richtig oder falsch, sondern mehr oder weniger sinnvoll
  - ein "sinnvolles" Clustering wird von den verschiedenen Verfahren auf der Grundlage von verschiedenen Annahmen (Heuristiken!) angestrebt
  - Überprüfung der Sinnhaftigkeit erfordert Fachwissen über die Datengrundlage

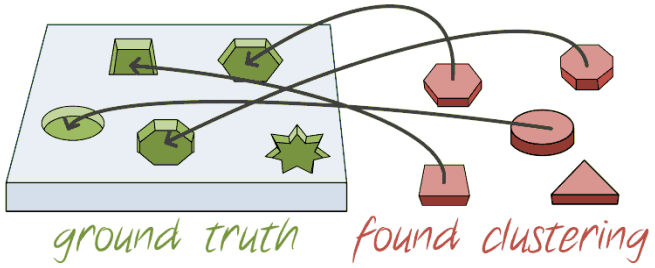
Evaluation of Clustering Results

General approaches:

- Evaluation based on **expert's** opinion
  - \+ may reveal new insight into the data
  - – very expensive, results are not comparable



- Evaluation based on **internal** measures
  - \+ no additional information needed
  - – approaches optimizing the evaluation criteria will always be preferred



- Evaluation based on **external** measures
  - \+ objective evaluation
  - – needs „ground truth"

  e.g., comparison of two clusterings

Evaluation based on internal measures

Idea:

- Make some assumptions on the characteristics of clusters and define an evaluation criteria thereof
- Measure how good the clusters of the clustering reflect those characteristics
- Example: Clusters
  - Minimize intra-cluster distances
  - Maximize inter-cluster distances
- Obvious downside: approaches optimizing the evaluation criteria will always be preferred (k-means-like are preffered over density-based in the above example)

Popular Measures

Given a clustering C= $(C_{1,\dots}, C_k)$ for Dataset DB

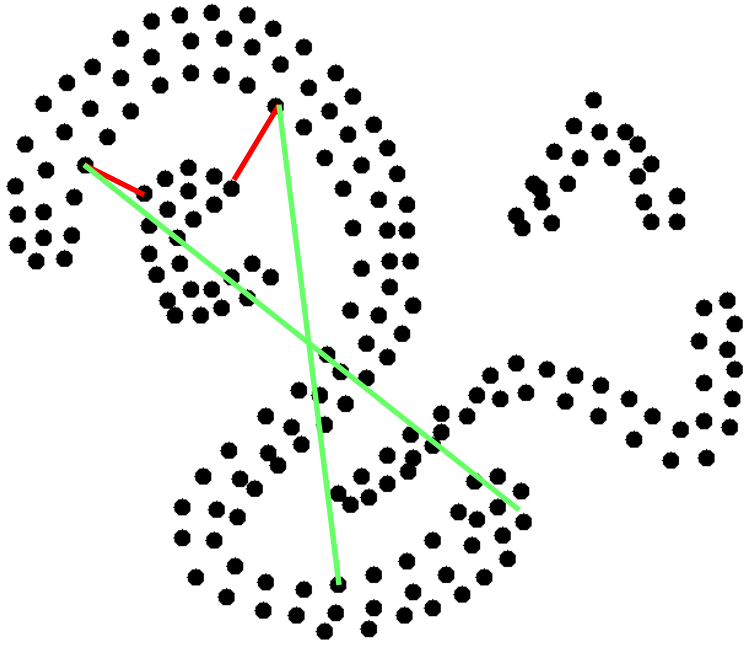
- Sum of square distances:

$$SSD(\mathrm{C}) = \frac{1}{|DB|} \sum_{C_i \in \mathrm{C}} \sum_{p \in C_i} (dist(p, \mu(C_i))^2$$

- Cohesion: measures the similarity of objects within a cluster
- Separation: measures the dissimilarity of one cluster to another one
- Silhouette Coefficient: combines cohesion and separation

Cohesion and Separation

- Suitable for globular cluster, but not for stretched clusters



Evaluation based on external measures

Idea:

Given

- a clustering $\mathcal{C} = (C_1, \ldots, C_k)$ and
- a "ground truth" $\mathcal{G} = (G_1, \ldots, G_l)$ for dataset DB

Measure the „similarity" between $\mathcal{C}$ and $\mathcal{G}$.

Evaluation based on external measures

Popular Measures

- Recall: $rec(C_i, G_j) = \frac{|C_i \cap G_j|}{|G_j|}$
- Precision: $prec(C_i, G_j) = \frac{|C_i \cap G_j|}{|C_i|}$
- F-Measure: $F(C_i, G_j) = \frac{2 * rec(C_i, G_j) * prec(C_i, G_j)}{rec(C_i, G_j) + prec(C_i, G_j)}$
- Purity (P): $P(\mathcal{C}, \mathcal{G}) = \sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|DB|} pur(C_i, \mathcal{G})$ $\quad pur(C_i, \mathcal{G}) = \max_{G_j \in \mathcal{G}} prec(C_i, G_j)$
- Rand Index: „normalized number of agreements “
- Jaccard Coefficient (JC)
- …

Evaluation based on external measures

Rand Index:

$$RI = \frac{a+d}{a+b+c+d}$$

a: #Paare in gleicher Klasse und gleichem Cluster

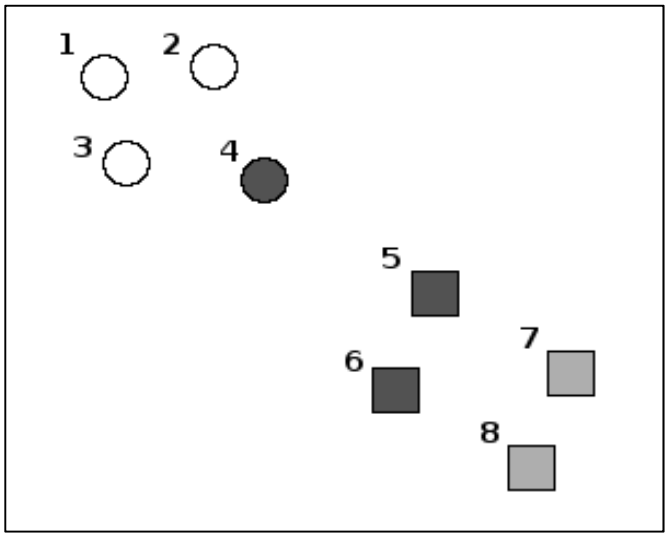
b: #Paare in gleicher Klasse aber versch. Clustern

c: #Paare in versch. Klassen aber gleichem Cluster

d: #Paare in versch. Klassen und versch. Cluster

Beispiel:

- Ground Truth: 2 Klassen (Kreise und Quadrate)
- 3 Cluster (Schwarz, Weiß, Grau)



a = 5; b = 7; c = 2; d = 14

RI = 5+14/(5+7+2+14) = 0.6785

Evaluation based on external measures

Jaccard Coefficient

$$Jc = \frac{a}{a+b+c}$$

a: #Paare in gleicher Klasse und gleichem Cluster

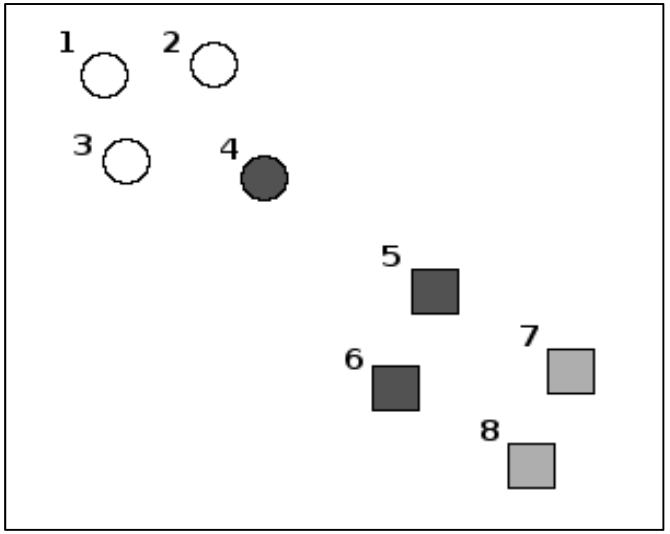
b: #Paare in gleicher Klasse aber versch. Clustern

c: #Paare in versch. Klassen aber gleichem Cluster

d: #Paare in versch. Klassen und versch. Cluster

Beispiel:

- 2 Klassen (Kreise und Quadrate)
- 3 Cluster (Schwarz, Weiß, Grau)



a = 5; b = 7; c = 2

Jc = 5/(5+7+2) = 0.3571

Evaluation based on external measures

Further methods:

- Mutual Entropy: $H(\mathcal{C}|\mathcal{G}) = -\sum_{C_i \in \mathcal{C}} p(C_i) \sum_{G_j \in \mathcal{G}} p(G_j|C_i) log\ p(G_j|C_i)$
  $= -\sum_{C_i \in \mathcal{C}} \frac{|C_i|}{|DB|} \sum_{G_j \in \mathcal{G}} \frac{|C_i \cap G_j|}{|C_i|} * \log_2 (\frac{|C_i \cap G_j|}{|C_i|})$

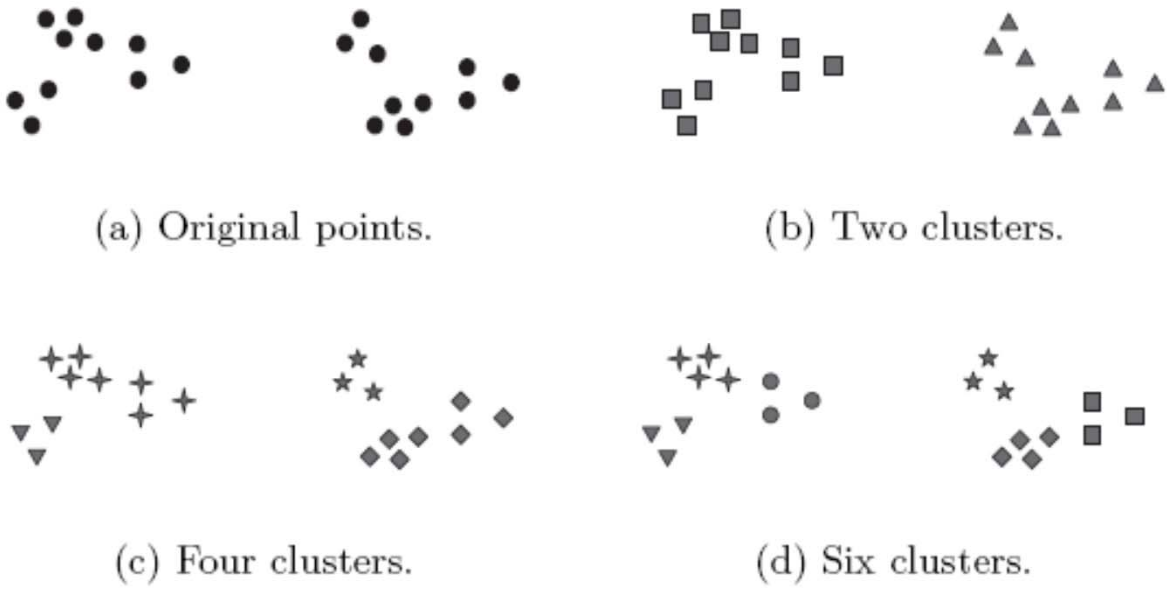
- Mutual Information: $I(\mathcal{C}, \mathcal{G}) = H(\mathcal{C}) - H(\mathcal{C}|\mathcal{G}) = H(\mathcal{G}) - H(\mathcal{G}|\mathcal{C})$

  where entropy $H(\mathcal{C}) = -\sum_{C_i \in \mathcal{C}} p(C_i) \cdot \log p(C_i) = -\sum_{C \in \mathcal{C}} \frac{|C_i|}{|DB|} \cdot \log \frac{|C_i|}{|DB|}$

- Normalized Mutual Information: $NMI(\mathcal{C}, \mathcal{G}) = \frac{I(\mathcal{C}, \mathcal{G})}{\sqrt{H(\mathcal{C}) H(\mathcal{G})}}$

Ambiguity of Clusterings

Different possibilities to cluster a set of objects



(a) Original points.

(b) Two clusters.

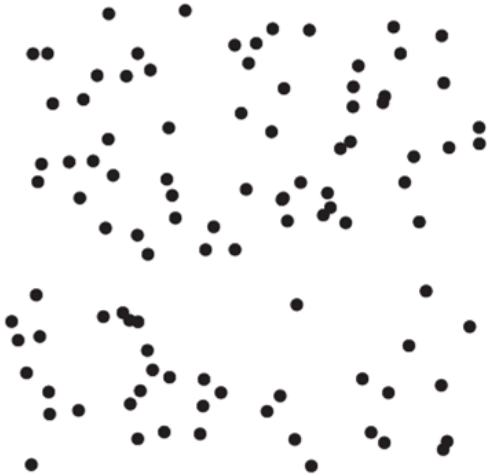
(c) Four clusters.

(d) Six clusters.

from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

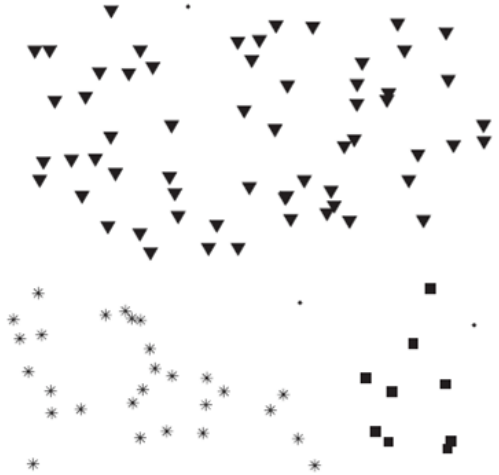
Ambiguity of Clusterings

Clusters on randomized data (equally distributed data)



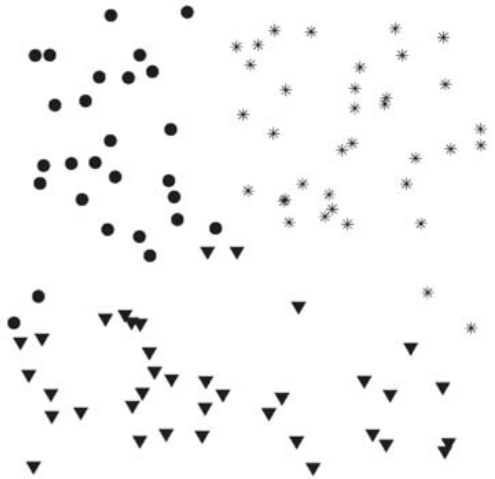
Data set
(100 equally distributed 2D points)

DBSCAN (3 Cluster)

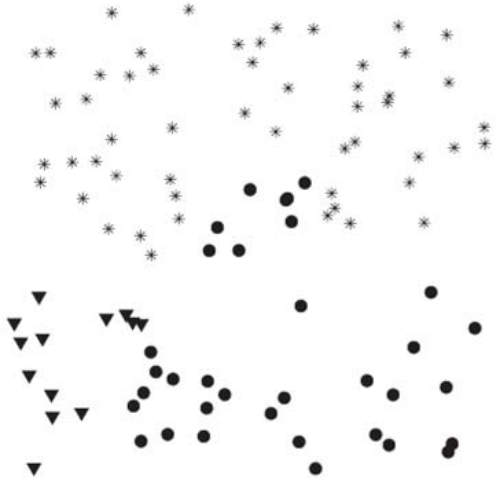


see: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

k-means
(3 Cluster)



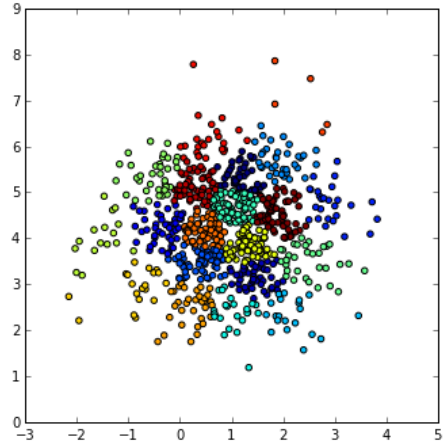
complete link
(3 Cluster)



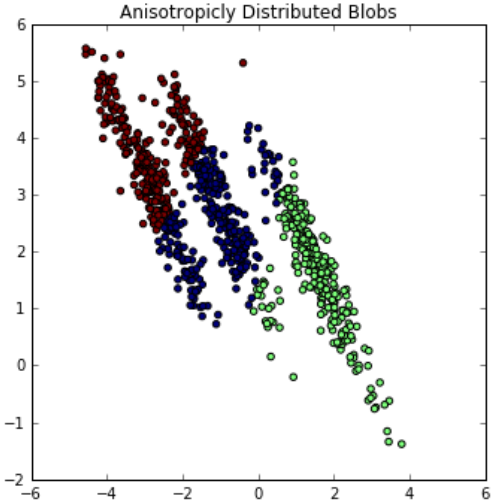
Ambiguity of Clusterings

- Kind of a philosophical problem:

  "What is a correct clustering?"
- Most approaches find clusters in every dataset, even in uniformly distributed object

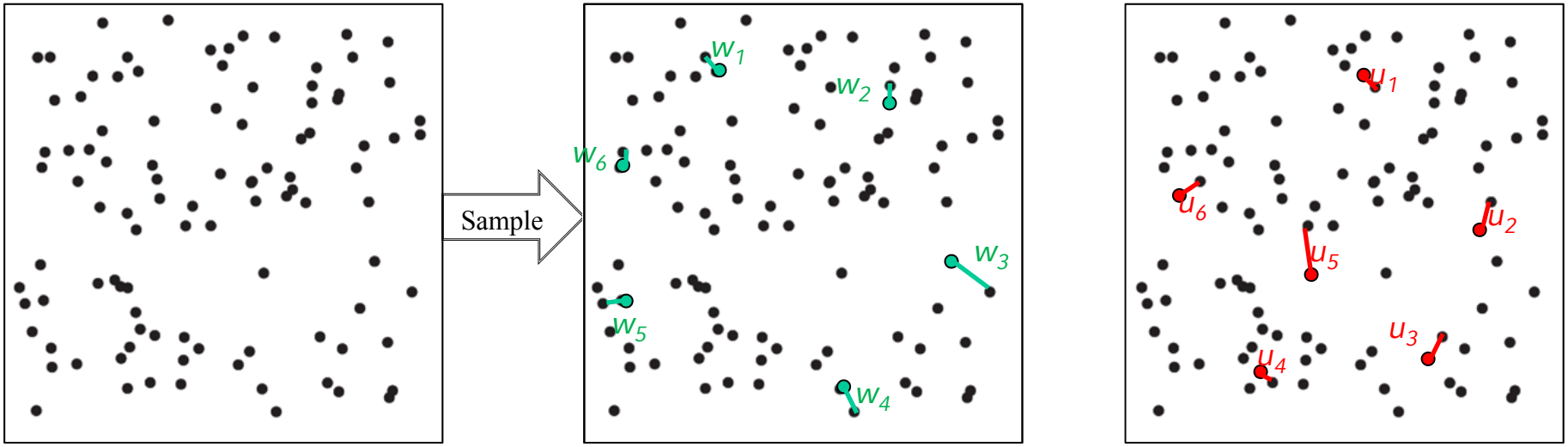


- Are there clusters?
  - Apply clustering algorithm
  - Check for reasonability of clusters
- Problem: No clusters found ⇏ no clusters existing
  - Maybe clusters exists only in certain models, but can not be found by used clustering approach



- Independent of clustering: Is there a data-given cluster tendency?

Hopkins statistics



dataset
(*n* objects)

Random selection
(*m* objects) $m \ll n$

m uniformly
distributed objects

w_i: distances of selected objects to the next neighbor in dataset

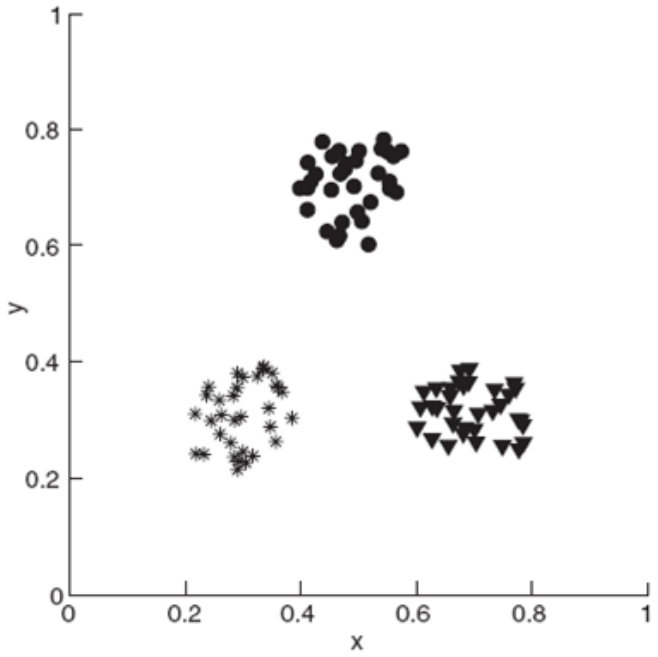
u_i: distances of uniformly distributed objects to the next neighbor in dataset

$$H = \frac{\sum_{i=1}^{m} u_i}{\sum_{i=1}^{m} u_i + \sum_{i=1}^{m} w_i}$$

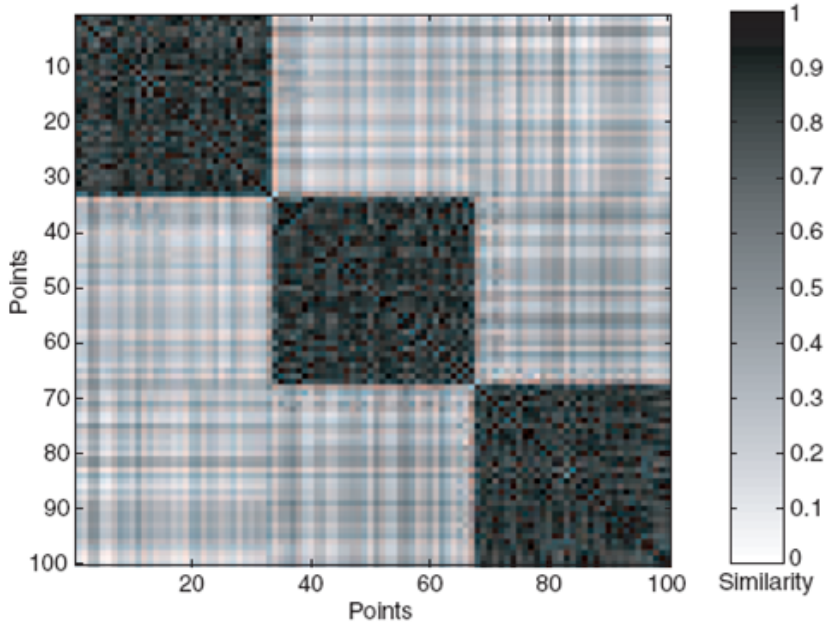
$0 \le H \le 1$

H ≈ 0 : data are very regular (e.g. on grid)
H ≈ 0,5 : data are uniformly distributed
H ≈ 1 : data are strongly clustered

Evaluating the Similarity Matrix



dataset
(well separated clusters)

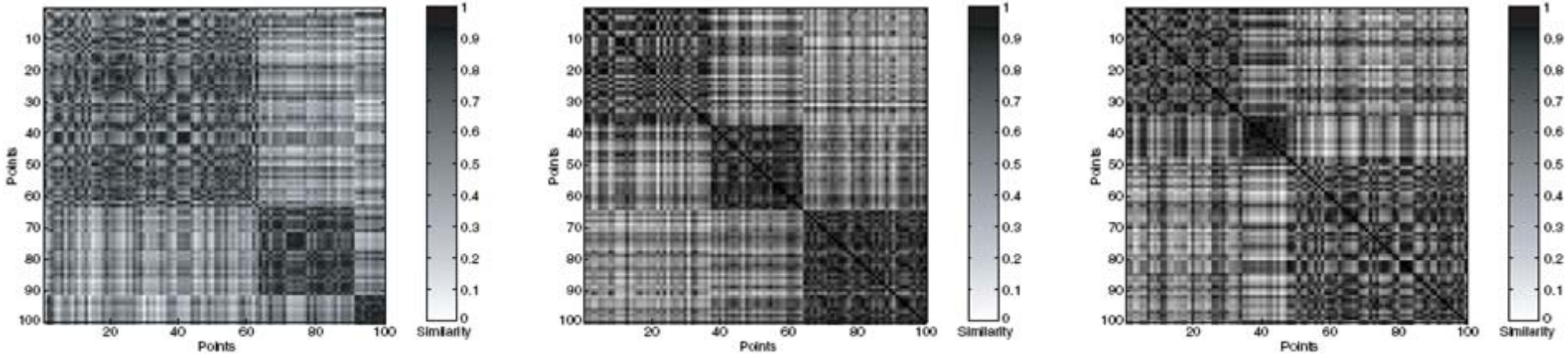


Similarity matrix
(sorted by k-means Cluster-labels)

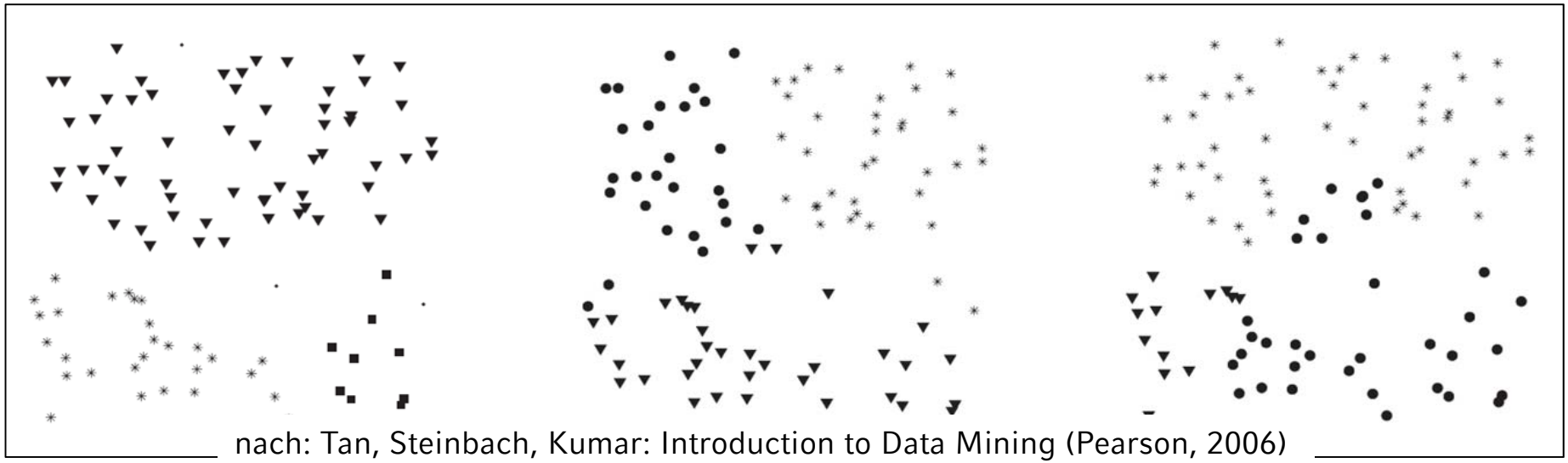
from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Evaluating the Similarity Matrix (cont'd)

similarity matrices differ for different clustering approaches



DBSCAN k-means complete link



nach: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Contents

1) Introduction to clustering
2) Partitioning Methods
   - K-Means
   - K-Medoid
   - Choice of parameters: Initialization, Silhouette coefficient
3) Expectation Maximization: a statistical approach
4) Density-based Methods: DBSCAN
5) Hierarchical Methods
   - Agglomerative and Divisive Hierarchical Clustering
   - Density-based hierarchical clustering: OPTICS
6) Evaluation of Clustering Results
7) Further Clustering Topics
   - Ensemble Clustering

Ensemble Clustering

Problem:

- Many differing cluster definitions
- Parameter choice usually highly influences the result

→ What is a ‚good' clustering?

Idea: Find a consensus solution (also ensemble clustering) that consolidates multiple clustering solutions.

Benefits of Ensemble Clustering:

- **Knowledge Reuse:** possibility to integrate the knowledge of multiple known, good clusterings
- **Improved quality**: often ensemble clustering leads to "better" results than its individual base solutions.
- **Improved robustness**: combining several clustering approaches with differing data modeling assumptions leads to an increased robustness across a wide range of datasets.
- **Model Selection:** novel approach for determining the final number of clusters
- **Distributed Clustering:** if data is inherently distributed (either feature-wise or object-wise) and each clusterer has only access to a subset of objects and/or features, ensemble methods can be used to compute a unifying result.

Ensemble Clustering – Basic Notions

Given: a set of L clusterings $\mathfrak{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ for dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathbb{R}^D$

Goal : find a consensus clustering $\mathcal{C}^*$

What exactly is a consensus clustering?

We can differentiate between 2 categories for ensemble clustering:

- Approaches based on pairwise similarity
  Idea: find a consensus clustering $\mathcal{C}^*$ for which the similarity function $\phi(\mathfrak{C}, \mathcal{C}^*) = \frac{1}{L}\sum_{l=1}^{L} \phi(\mathcal{C}_l, \mathcal{C}^*)$ is maximal.

  basically our external evaluation measures, which compare two clusterings

- Probabilistic approaches :
  Assume that the L labels for the objects $\mathbf{x}_i \in \mathbf{X}$ follow a certain distribution

→ We will present one exemplary approach for both categories in the following

Ensemble Clustering – Similarity-based Approaches

Given: a set of L clusterings $\mathfrak{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ for dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathbb{R}^D$

- Goal : find a consensus clustering $\mathcal{C}^*$ for which the similarity function $\phi(\mathfrak{C}, \mathcal{C}^*) = \frac{1}{L}\sum_{l=1}^{L} \phi(\mathcal{C}_l, \mathcal{C}^*)$ is maximal.
  - Popular choices for ϕ in the literature:
    - **Pair counting-based measures:** Rand Index (RI), Adjusted RI, Probabilistic RI
    - **Information theoretic measures:** Mutual Information (I), Normalized Mutual Information (NMI), Variation of Information (VI)

Problem: the above objective is intractable

Solutions:

- Methods based on the co-association matrix (related to RI)
- Methods using cluster labels without co-association matrix (often related to NMI)
  - Mostly graph partitioning
  - Cumulative voting

Ensemble Clust. – Approaches based on Co-Association

Given: a set of L clusterings $\mathfrak{C} = \{\mathcal{C}_1, \dots, \mathcal{C}_L\}$ for dataset $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_N\} \in \mathbb{R}^D$

The co-association matrix $\mathbf{S}^{(\mathfrak{C})}$ is an $N \times N$ matrix representing the label similarity of object pairs: $s_{i,j}^{(\mathfrak{C})} = \sum_{l=1}^{L} \delta\left(C(\mathcal{C}_l, \mathbf{x}_i), C(\mathcal{C}_l, \mathbf{x}_j)\right)$

where $\delta(a,b) = \begin{cases} 1 & if\ a = b \\ 0 & else \end{cases}$

cluster label of $\mathbf{x}_i, \mathbf{x}_j$ in clustering $\mathcal{C}_l$

Based on the similarity matrix defined by $\mathbf{S}^{(\mathfrak{C})}$ traditional clustering approaches can be used

Often $\mathbf{S}^{(\mathfrak{C})}$ is interpreted as weighted adjacency matrix, such that methods for graph partitioning can be applied.

In [Mirkin'96] a connection of consensus clustering based on the co-association matrix and the optimization of the pairwise similarity based on the Rand Index ($\mathcal{C}^{best} = argmax_{\mathcal{C}^*} \left\{\frac{1}{L}\sum_{\mathcal{C}_l \in \mathfrak{C}} RI(\mathcal{C}_l, \mathcal{C}^*)\right\}$) has been proven.

[Mirkin'96] B. Mirkin: *Mathematical Classification and Clustering*. Kluwer, 1996.

Ensemble Clust. – Approaches based on Cluster Labels

- Consensus clustering $\mathcal{C}^*$ for which $\frac{1}{L}\sum_{\mathcal{C}_l \in \mathfrak{C}} \phi(\mathcal{C}_l, \mathcal{C}^*)$ is maximal

→ Information theoretic approach: choose ϕ as mutual information (I), normalized mutual information (NMI), information bottleneck (IB),…

- Problem: Usually a hard optimization problem

→ Solution 1: Use meaningful optimization approaches (e.g. gradient descent) or heuristics to approximate the best clustering solution (e.g. [SG02])

→ Solution 2: Use a similar but solvable objective (e.g. [TJP03])

- Idea: use as objective $\mathcal{C}_{best} = argmax_{\mathcal{C}^*}\left\{\frac{1}{L}\sum_{\mathcal{C}_l \in \mathfrak{C}} I^s(\mathcal{C}_l, \mathcal{C}^*)\right\}$
  where I^s is the mutual information based on the generalized entropy of degree s: $H^s(X) = (2^{1-s}-1)^{-1}\sum_{x_i \in X}(p_i^s - 1)$

→ For s = 2, $I^s(\mathcal{C}_l, \mathcal{C}^*)$ is equal to the category utility function whose maximization is proven to be equivalent to the minimization of the square-error clustering criterion

→ Thus apply a simple label transformation and use e.g. K-Means

[SG02] A. Strehl, J. Ghosh: *Cluster ensembles - a knowledge reuse framework for combining multiple partitions.* Journal of Machine Learning Research, 3, 2002, pp. 583-617.
[TJP03]A. Topchy, A.K. Jain, W. Punch. *Combining multiple weak clusterings.* In ICDM, pages 331-339, 2003.

Ensemble Clustering – A Probabilistic Approach

Assumption 1: all clusterings $\mathcal{C}_l \in \mathfrak{C}$ are partitionings of the dataset $\mathbf{X}$.

Assumption 2: there are K^* consensus clusters

cluster label of x_n in clustering $\mathcal{C}_l$

The dataset $\mathbf{X}$ is represented by the set

$$\mathbf{Y} = \{\boldsymbol{y}_n \in \mathbb{N}_0^L | \exists \mathbf{x}_n \in \mathbf{X}. \forall \mathcal{C}_l \in \mathfrak{C}. y_{nl} = C(\mathcal{C}_l, \mathbf{x}_n)\}$$

we have a new feature Space $\mathbb{N}_0^L$, where the l^{th} feature represents the cluster labels from partition $\mathcal{C}_l$

Assumption 3: the dataset $\mathbf{Y}$ (labels of base clusterings) follow a multivariate mixture distribution:

conditional independence assumptions for $\mathcal{C}_l \in \mathfrak{C}$

$$P(\mathbf{Y}|\boldsymbol{\Theta}) = \prod_{n=1}^{N} \sum_{k=1}^{K^*} \alpha_k P_k(\boldsymbol{y}_n|\boldsymbol{\theta}_k) = \prod_{n=1}^{N} \sum_{k=1}^{K^*} \alpha_k \prod_{l=1}^{L} P_{kl}(y_{nl}|\boldsymbol{\theta}_{kl})$$

$P_{kl}(y_{nl}|\boldsymbol{\theta}_{kl}) \sim M\left(1, \left(p_{k,l,1}, \ldots, p_{k,l,|\mathcal{C}_l|}\right)\right)$ follows a $|\mathcal{C}_l|$-dimensional multinomial distribution: $P_{kl}(y_{nl}|\boldsymbol{\theta}_{kl}) = \prod_{k'=1}^{|\mathcal{C}_l|} p_{k,l,k'}^{\delta(y_{nl},k')}$

therefore: $\boldsymbol{\theta}_{kl} = (p_{k,l,1}, \ldots, p_{k,l,|\mathcal{C}_l|})$

Goal: find the parameters $\boldsymbol{\Theta} = (\alpha_1, \boldsymbol{\theta}_1, \ldots, \alpha_{K^*}, \boldsymbol{\theta}_{K^*})$ such that the likelihood $P(\mathbf{Y}|\boldsymbol{\Theta})$ is maximized

Solution: optimizing the parameters via the EM approach. *(details omitted)*

Presented approach: Topchy, Jain, Punch: *A mixture model for clustering ensembles*. In ICDM, pp. 379-390, 2004.
Later extensions: H. Wang, H. Shan, A. Banerjee: *Bayesian cluster ensembles*. In ICDM, pp. 211-222, 2009.
P. Wang, C. Domeniconi, K. Laskey: *Nonparametric Bayesian clustering ensembles*. In PKDD, pp. 435-450, 2010.

Contents

1) Introduction to clustering
2) Partitioning Methods
   - K-Means
   - K-Medoid
   - Choice of parameters: Initialization, Silhouette coefficient
3) Expectation Maximization: a statistical approach
4) Density-based Methods: DBSCAN
5) Hierarchical Methods
   - Agglomerative and Divisive Hierarchical Clustering
   - Density-based hierarchical clustering: OPTICS
6) Evaluation of Clustering Results
7) Further Clustering Topics
   - Ensemble Clustering
   - Discussion: an alternative view on DBSCAN

Reconsider DBSCAN algorithm

- Standard DBSCAN evaluation is based on recursive database traversal.
- Böhm et al. (2000) observed that DBSCAN, among other clustering algorithms, may be efficiently built on top of similarity join operations.

Similarity joins

- An ε-similarity join yields all pairs of ε-similar objects from two data sets P, Q:

$$P \bowtie_\varepsilon Q = \{(p,q) | \, p \in P \wedge q \in Q \wedge dist(p,q) \leq \varepsilon\}$$

  - SQL-Query: SELECT * FROM P, Q WHERE $dist(\mathrm{P},\mathrm{Q}) \leq \varepsilon$

- An ε-similarity self join yields all pairs of ε-similar objects from a database DB:

$$DB \bowtie_\varepsilon DB = \{(p,q) | \, p \in DB \wedge q \in DB \wedge dist(p,q) \leq \varepsilon\}$$

  - SQL-Query: SELECT * FROM DB p, DB q WHERE $dist(p,q) \leq \varepsilon$

Böhm C., Braunmüller, B., Breunig M., Kriegel H.-P.: High performance clustering based on the similarity join. CIKM 2000: 298-305.

ε-Similarity self join:

$$DB \bowtie_\varepsilon DB = \{(p,q) | \, p \in DB \wedge q \in DB \wedge dist(p,q) \leq \varepsilon\}$$

Relation „directly ε, *MinPts*-density reachable“ may be expressed in terms of an ε-similarity self join:

$$ddr_{\varepsilon,\mu} = \{(p,q) | \, p \text{ is } \varepsilon,\mu\text{–core point} \wedge q \in N_\varepsilon(p)\}$$

$$= \{(p,q) | p,q \in DB \wedge dist(p,q) \leq \varepsilon \wedge \exists_{\geq \mu} q' \in DB : dist(p,q') \leq \varepsilon\}$$

$$= \{(p,q) | (p,q) \in DB \bowtie_\varepsilon DB \wedge \exists_{\geq \mu} q' : (p,q') \in DB \bowtie_\varepsilon DB\}$$

$$= \sigma_{|\pi_p(DB \bowtie_\varepsilon DB)| \geq \mu}(DB \bowtie_\varepsilon DB) =: DB \bowtie_{\varepsilon,\mu} DB$$

- SQL-Query: SELECT * FROM DB p, DB q WHERE $dist(p,q) \leq \varepsilon$ GROUP BY p.id HAVING $\text{count}(p.id) \geq \mu$
- Remark: $DB \bowtie_\varepsilon DB$ is a symmetric relation, $ddr_{\varepsilon,\mu} = DB \bowtie_{\varepsilon,\mu} DB$ is not.

DBSCAN then computes the connected components within $DB \bowtie_{\varepsilon,\mu} DB$.

Clustering – Summary

Partitioning Methods: K-Means, K-Medoid, K-Mode, K-Median

Probabilistic Model-Based Clusters: Expectation Maximization

Density-based Methods: DBSCAN

Hierarchical Methods

- Agglomerative and Divisive Hierarchical Clustering
- Density-based hierarchical clustering: OPTICS

Evaluation of Clustering Results

- Evaluation based on an expert's knowledge; internal evaluation measures; external evaluation measures

Further Clustering Topics

- Ensemble Clustering: finding a consensus clustering agreeing with multiple base clustering and its advantages
  - co-assocaition matrix, information theoretic approaches, probabilistic approaches
- Discussion of DBSCAN as Join operation

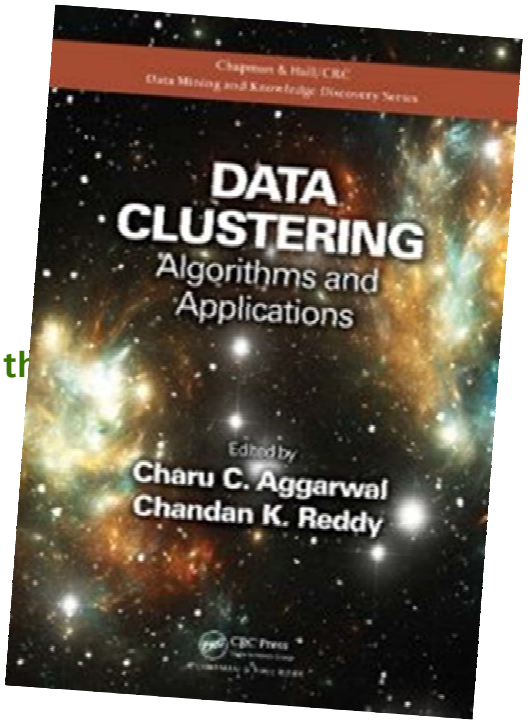
Further Literature for Clustering

Data Clustering – Algorithms and Applications

by C. C. Aggarwal and C. K. Reddy

published in August 2013 by Chapman & Hall/CRC

- Link to publisher
- Link to extract of the book



Contents:

this course

- **Probabilistic Models for Clustering**
- **A Survey of Partitional and Hierarchical Clustering Algorith**
- **Density-Based Clustering**
- **Clustering Categorical Data**
- **Cluster Ensembles: Theory and Applications**
- **Clustering Validation Measures**
- Feature Selection for Clustering
- Clustering High-Dimensional Data
- Spectral Clustering
- Network Clustering
- A Survey of Stream Clustering Algorithms
- Alternative Clustering Analysis: A Review
- Time-Series Data Clustering
- Clustering Multimedia Data
- …
- And many many more (e.g. Big Data Clustering, Document Clustering, …)