Lecture Notes

Foundations of Data Analysis

summer term 2020

Lecture: Univ.-Prof. Dr. Claudia Plant
Assistance: Lukas Miklautz
Tutor: Claus Hofmann

Evaluation of Unsupervised Methods

Overview:

- **1 Evaluation of clusters and outliers**
- **2 Evaluation of clustering-algorithms**
- **3 Evaluation of outlier-detection-algorithms**

Evaluation of clusters and outliers

- Evaluation of clusters/outliers:
  - Applying a technique onto a concrete problem (a specific data set one wants to analyze)

    Remember:

    Knowledge Discovery in Databases (*KDD*) is the process of (un-)supervised extraction of knowledge from databases that is *new*, *valid* and (potentially) *useful*.
  - Question: What to do with this new knowledge?

- Evaluation of clustering-/outlier-detection-*algorithms*
  - Not necessarily new knowledge, but verifiable
  - Application to data that is already well known
  - Application onto artificial data whose structure is known by design
  - Question: Are properties/structures found that the algorithm should find according to its model? Is it better than other algorithms?
  - Verifiability alone is questionable!

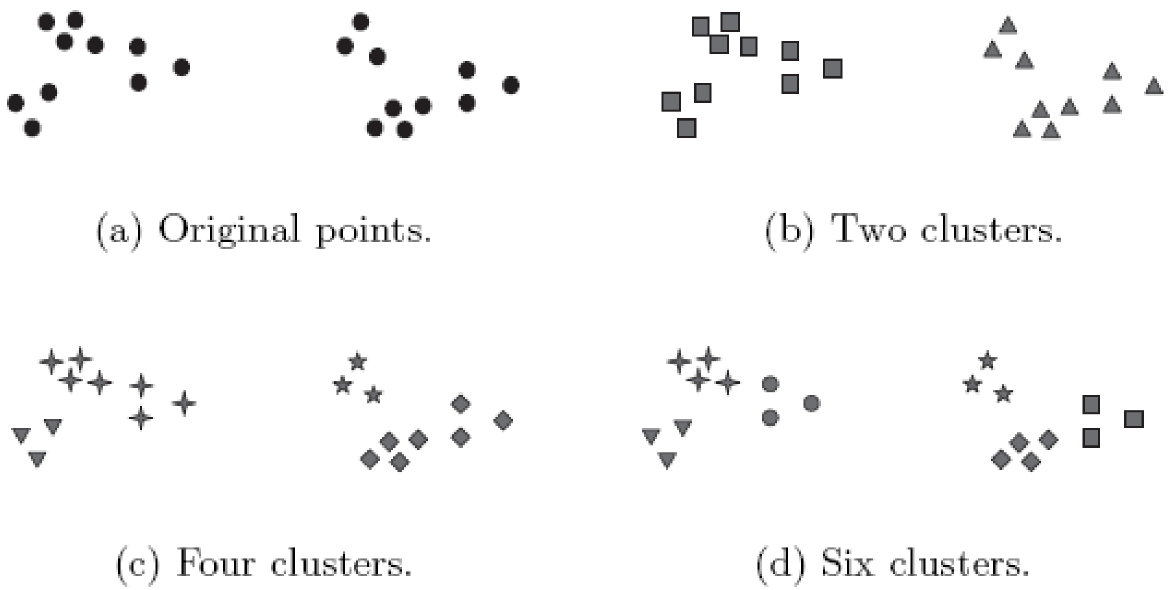
Evaluation of clusters and outliers

In principal: Clustering is unsupervised!

- Clustering is not right or wrong, but more or less useful or valid
- A "meaningful" clustering is aimed for on the basis of different assumptions (heuristics!) by the various algorithms
- Meaningful verification requires expert knowledge of the data set

Evaluation of clusters and outliers

- Different possibilities to cluster a data set



(a) Original points.

(b) Two clusters.

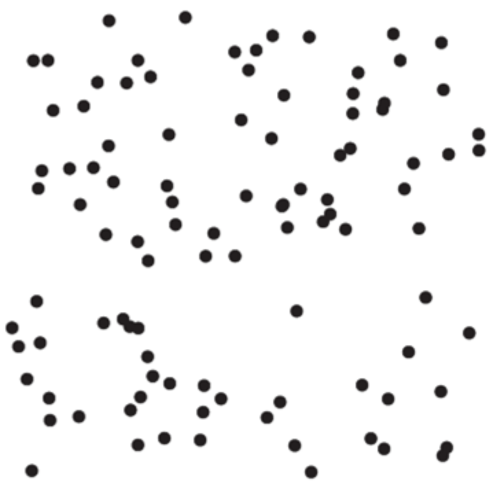
(c) Four clusters.

(d) Six clusters.

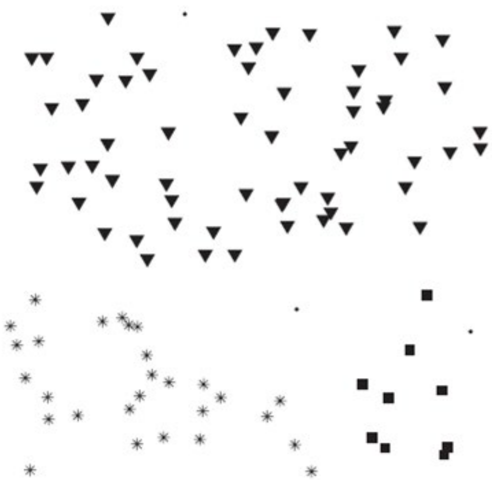
from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Evaluation of clusters and outliers

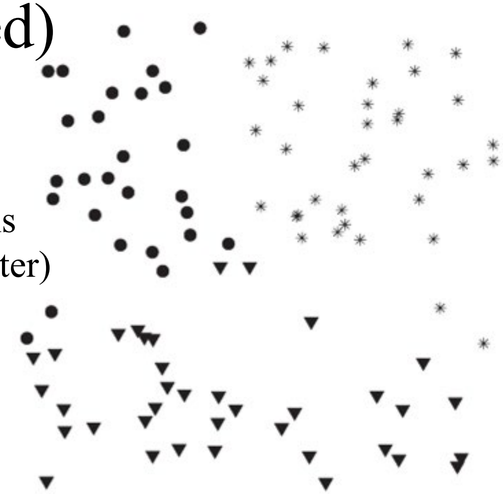
- Cluster results for random data (equally distributed)



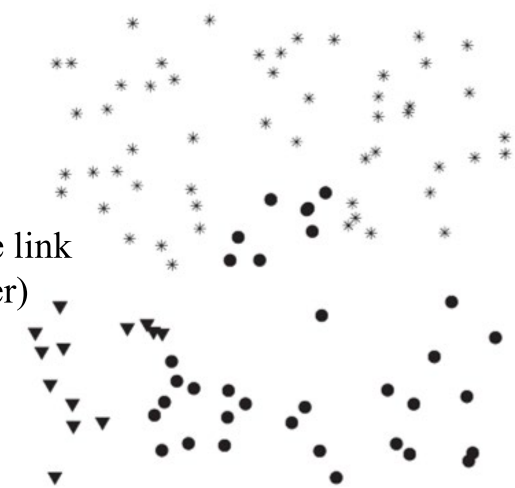
data set
(100 uniform distributed 2D data points)



DBSCAN (3 Cluster)



k-means
(3 Cluster)



complete link
(3 Cluster)

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

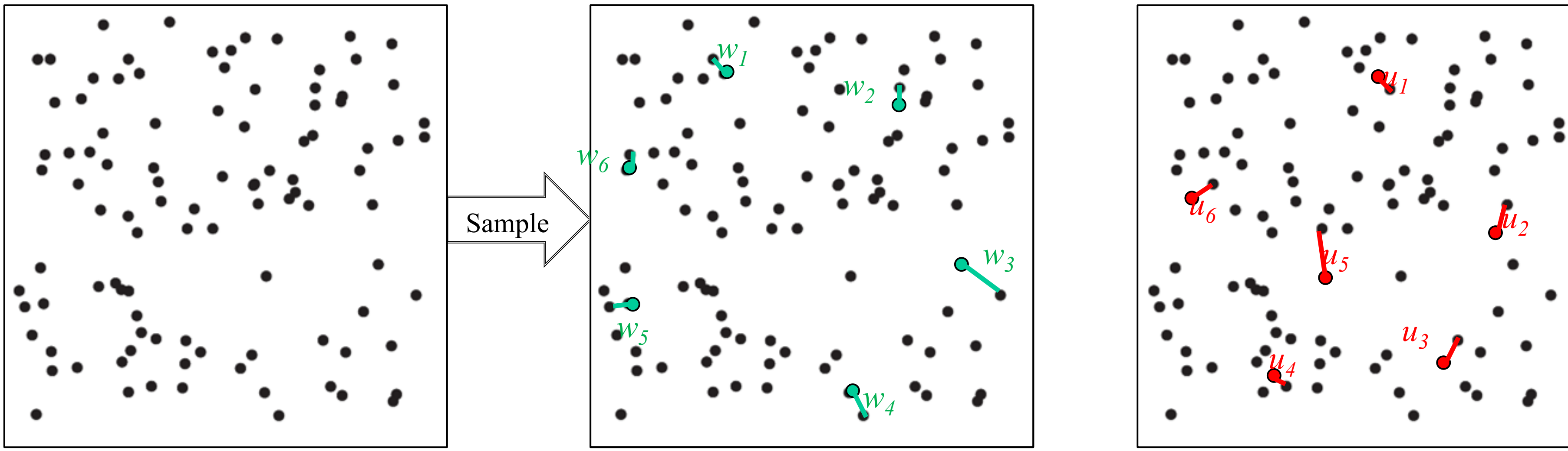
Evaluation of clusters and outliers

- Are there cluster in a data set?
- Many algorithms find clusters in every data set, regardless of whether they are there or not.
- Test whether clusters are in a data set:
  - Apply a cluster algorithm
  - Test if at least some of the found clusters make sense

- Problem: Negative results don't say anything
  - There may be clusters corresponding to another model (not found by the method used)

Evaluation of clusters and outliers

- Hopkins Statistics for clusters



Data
(n Points)

Random data points
(m Points) $m << n$

m uniform distributed
data points

- w_i: Distances of the selected points to their nearest neighbor in the original data set
- u_i: Distances of the uniform distributed points to their nearest neighbor in the original data set

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

$0 \leq H \leq 1$

$H \approx 0$: data very regular (e.g. on a grid)
$H \approx 0{,}5$: data is uniform
$H \approx 1$: data has strong cluster structure

Evaluation of clusters and outliers

- Evaluation of outliers:
- Strong possibility detected by an outlier detection method

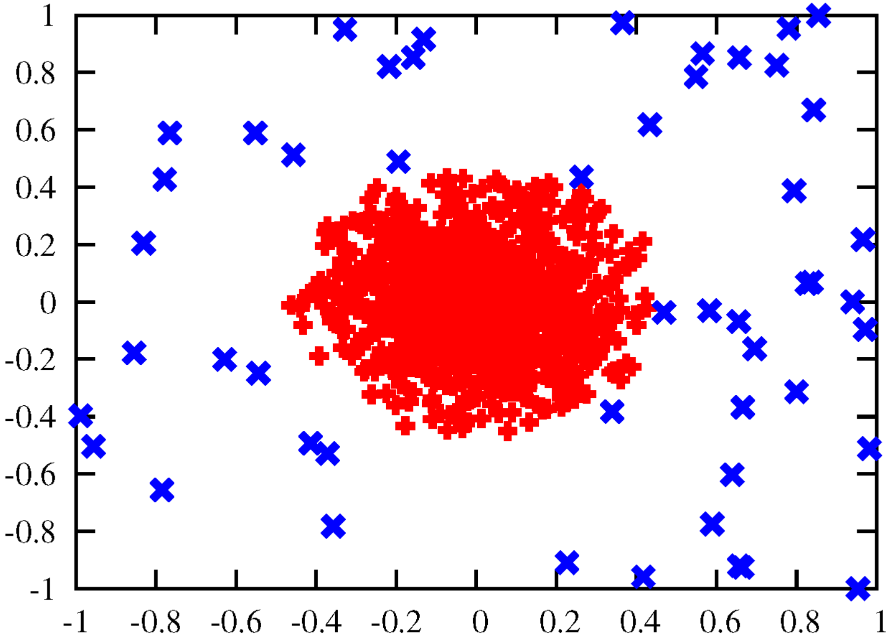
$\Rightarrow$ the object could be an outlier

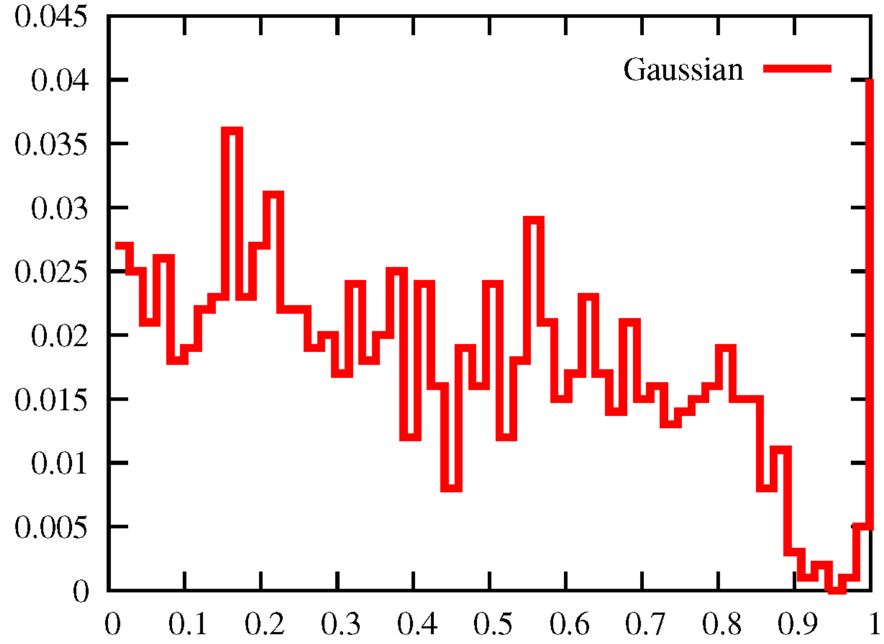
- Difficulty in defining what an outlier actually is:
  - "an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data" (Barnett, Lewis 1994)
  - "an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism“ (Hawkins 1980)

- Evaluation (or decision) requires knowledge of the data basis

Evaluation of clusters and outliers

- Outlier Scores:

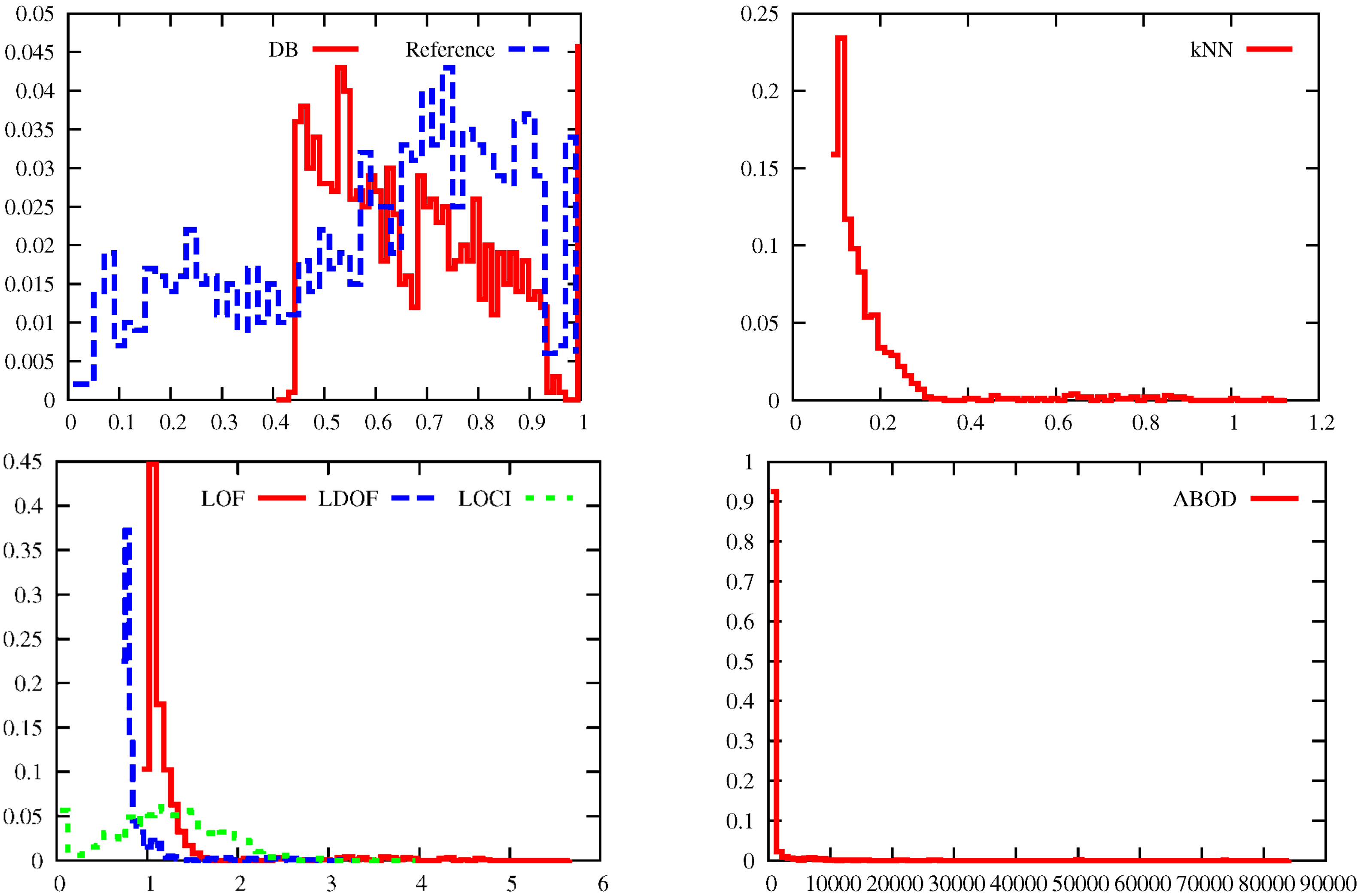




- *Hawkins (1980): "a sample containing outliers would show up such characteristics as large gaps between 'outlying' and 'inlying' observations and the deviation between outliers and the group of inliers, as measured on some suitably standardized scale"*
- Often a good ranking, but no "*large gap*": No definite descision outlier/inlier

Evaluation of clusters and outliers

universität wien
Fakultät für Informatik

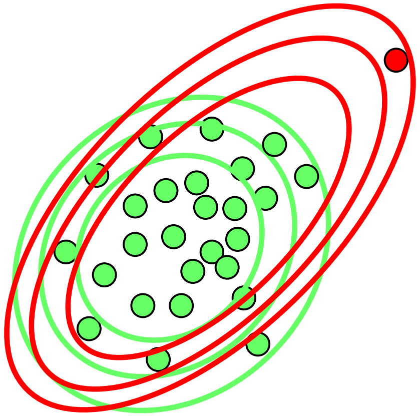


Evaluation of clusters and outliers

- *false positive* and *false negative* outliers:

  masking and *swamping* Effekt:

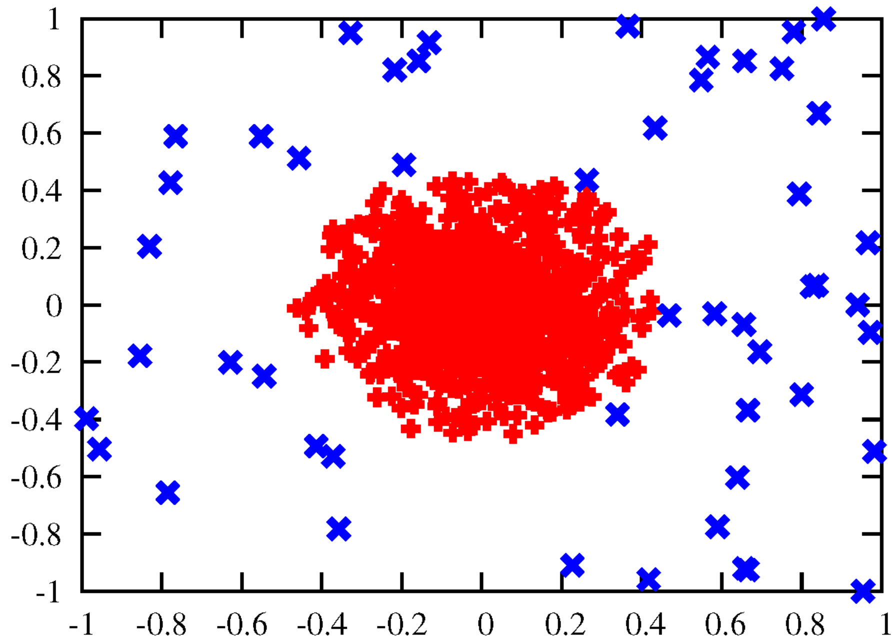
- outlier, which are taken into account in modelling, influence the model
- *masking*: the model is so strongly influenced that the outlier is explained by the model, i.e. it is masked
- *swamping*: due to the distortion of the model, inlier are no longer well explained by the model, are suspected to be Outlier



Evaluation of clusters and outliers

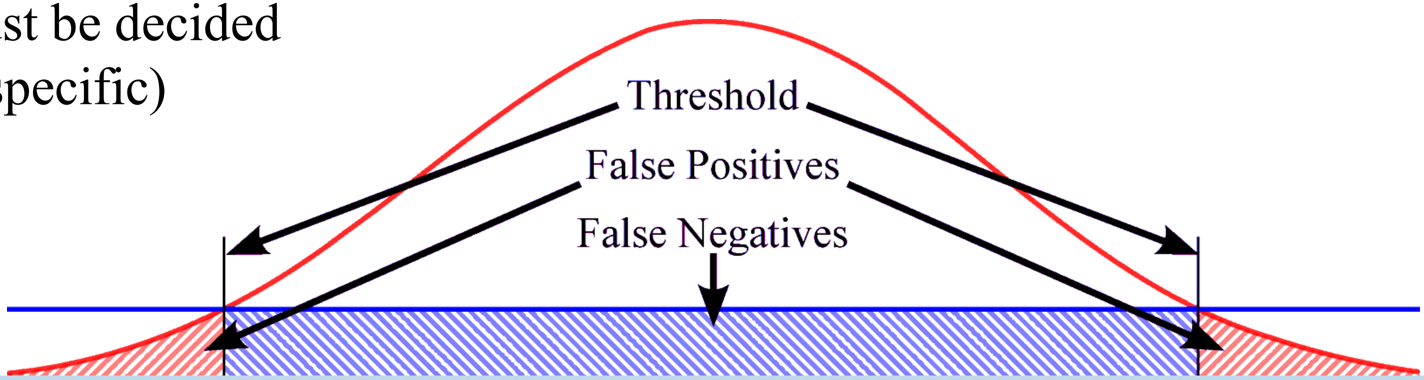
universität wien
Fakultät für Informatik

necessary levels of *false positive* and *false negative*:



- Objects, actually generated by a different process remain undetected because they match the distribution of normal objects very well
- normal objects in the tail of the "normal" distribution appear as outlier

→ actual outlier must be decided manually (domain-specific)



Alternatives to evaluation:

- "internal evaluation" (≈ unsupervised)
  - inner meaning (how well does the found model explain the data?)
  - Cohesion/Separation (examples: TD2, silhouette coefficient)
  - Similarity matrix (correlation, visualization)
  - Prerequisite: the procedure is principally appropriate for the problem

Evaluation of clustering-algorithms

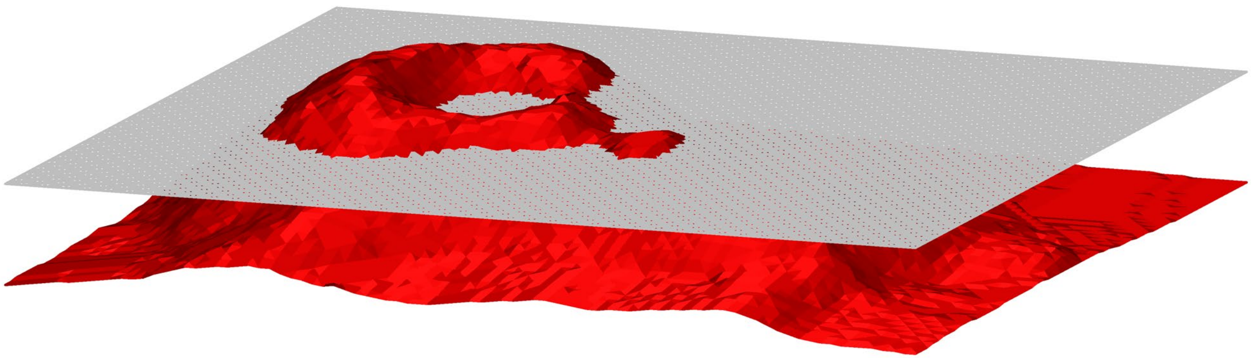
universität wien
Fakultät für Informatik

Alternatives to evaluation

- "external evaluation" (≈ supervised)
  - Checking the results independently of the algorithm on data with known properties
  - Example: Rediscover known classes
  - Application on data from the same area then probably makes sense
  - Problem: Discovery of new knowledge is more difficult

7.2 Evaluation of clustering-algorithms

- Internal Evaluation - Basic problem:
  - Adequacy of the algorithm
  - Deciding *before* use of the algorithm depending on properties of the algorithm and expected characteristics of the data and clusters:
- Type of clustering
  - e.g. biological taxonomy or geological topography: hierarchical or density-based
  - Clustering for compression: partitioning



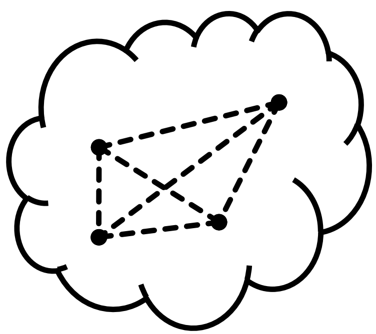
Evaluation of clustering-algorithms

Adequacy of the algorithm

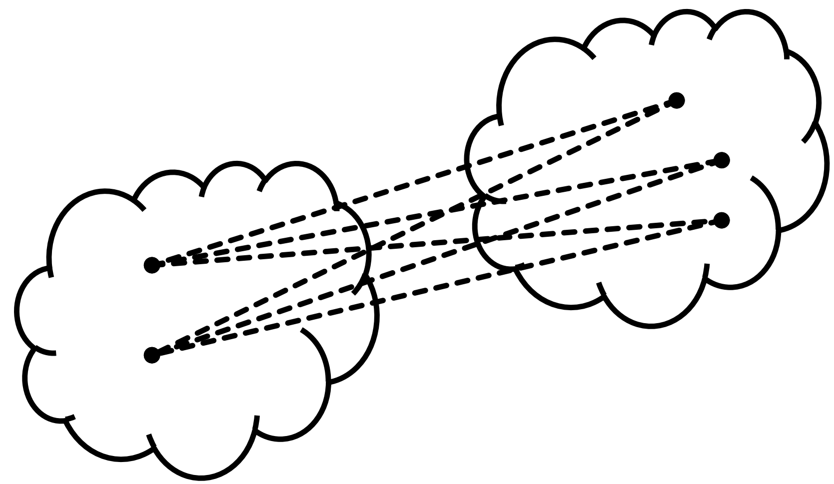
- Characteristics of the data set/attributes
  - e.g. k-means etc.: Mean value and variance must be calculated and interpreted meaningfully for the data
  - other (e.g. hierarchical procedures): the nature of the data is less important as long as a similarity matrix can be generated
- Noise/Outlier
  - EM/k-means: possibly strong model distortion due to outliers
  - density-based: stronger robustness against outliers

Cohesion and separation:

- Cohesion: to what extent is the cluster connected?
- Separation: How well is the cluster separated from other clusters?



Cohesion:
Small distances within the cluster



Separation :
Large distances between clusters

- Validity measure: suitable combination of cohesion and separation

Evaluation of clustering-algorithms

- Validity measure for a set of *k* clusters, $C_1,\ldots,C_k$, weight w_i of cluster C_i:

$$TotalValidity = \sum_{i=1}^{k} w_i \cdot Validity(C_i)$$

- Weight, e.g dependent on size of clusters
- With a given measure for proximity, e.g. distance function oder sz.B. Distanzfunktion, Ähnlichkeitsfunktion, ausgedrückt:

$$cohesion(C_i) = \sum_{\substack{x \in C_i \\ y \in C_i}} proximity(x, y)$$

$$separation(C_i, C_j) = \sum_{\substack{x \in C_i \\ y \in C_j}} proximity(x, y)$$

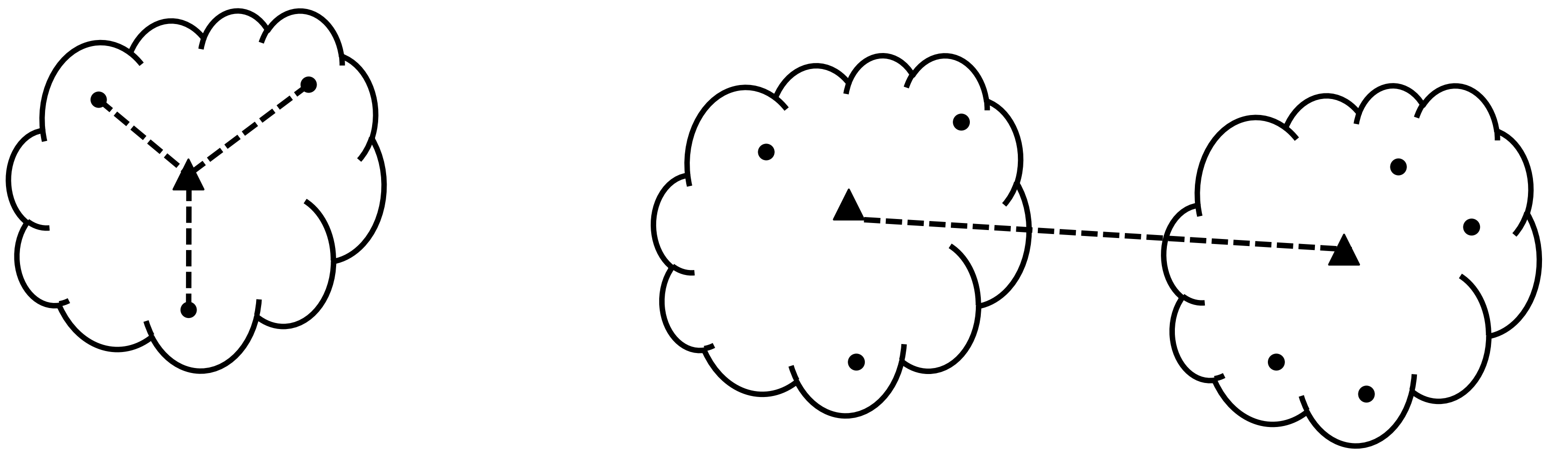
Evaluation of clustering-algorithms

universität wien
Fakultät für Informatik

Simple Example: ▲c_i für Cluster C_i:

$$cohesion(C_i) = \sum_{x \in C_i} proximity(x, c_i)$$

$$separation(C_i, C_j) = proximity(c_i, c_j)$$



Example: Cohesion

- *Centroid* μ_C: Mean of all datapoints in cluster *C*
- *Measure for the cost* (Cohesion) *of a cluster C*

$$TD^2(C) = \sum_{p \in C} dist(p, \mu_C)^2$$

- *Measure for the cost* (Cohesion) *of a clustering*

$$TD^2 = \sum_{i=1}^{k} TD^2(C_i)$$

Example: Silhouette coefficient

- Problem with many measures: Dependence on the number of clusters
- for k-means and k-medoid: TD2 and TD decrease monotonically with increasing k
- for EM: E rises monotonically with increasing k

- Silhouette coefficient
  - $a(o)$: Distance of a object o to representant of its cluster
  - $b(o)$: Distance to representant of „second-closest" clusters
  - Silhouette $s(o)$ of o:

$$s(o) = \frac{b(o) - a(o)}{\max\{a(o), b(o)\}}$$

$$-1 \leq s(o) \leq +1$$

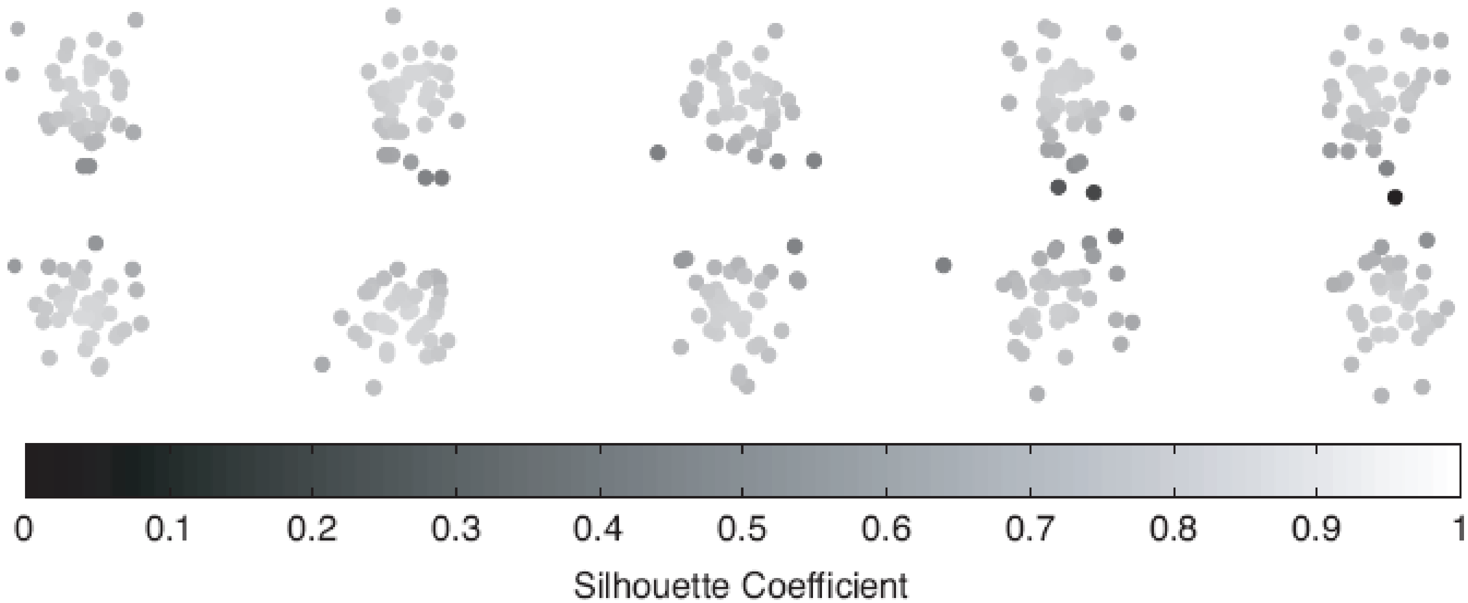
$$s(o) \approx -1 \;/\; 0 \;/\; +1 : \text{bad/ neither / good}$$

  - Silhouette coefficient s_C of a clustering: average silhouette of all objects
  - Interpretation of silhouette coefficient

$s_C > 0{,}7$: strong structure

$s_C > 0{,}5$: useful structure

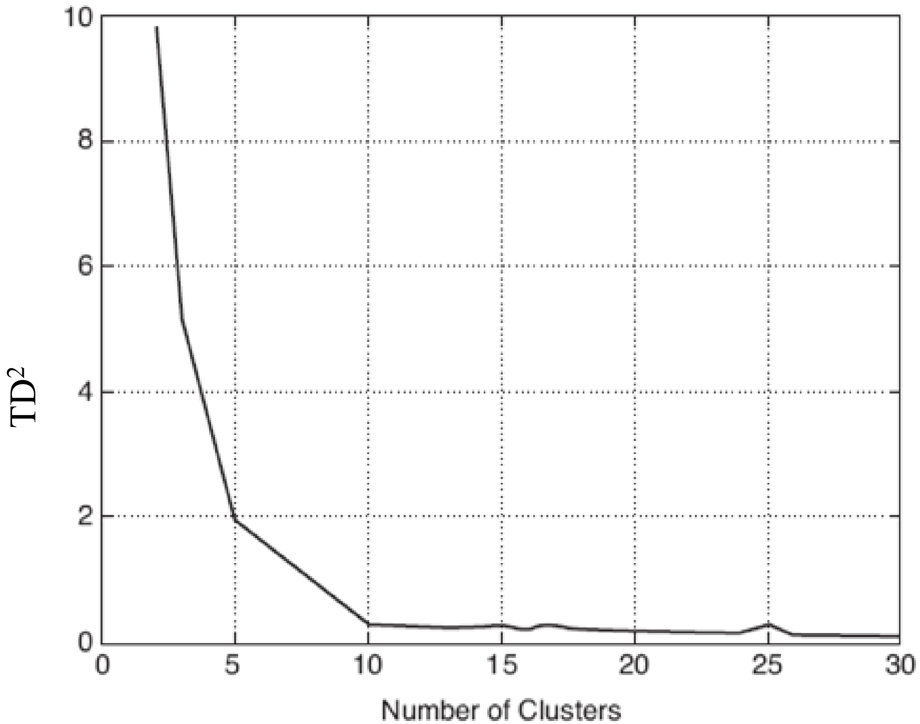
Silhouette-Coefficient for points in 10 clusters

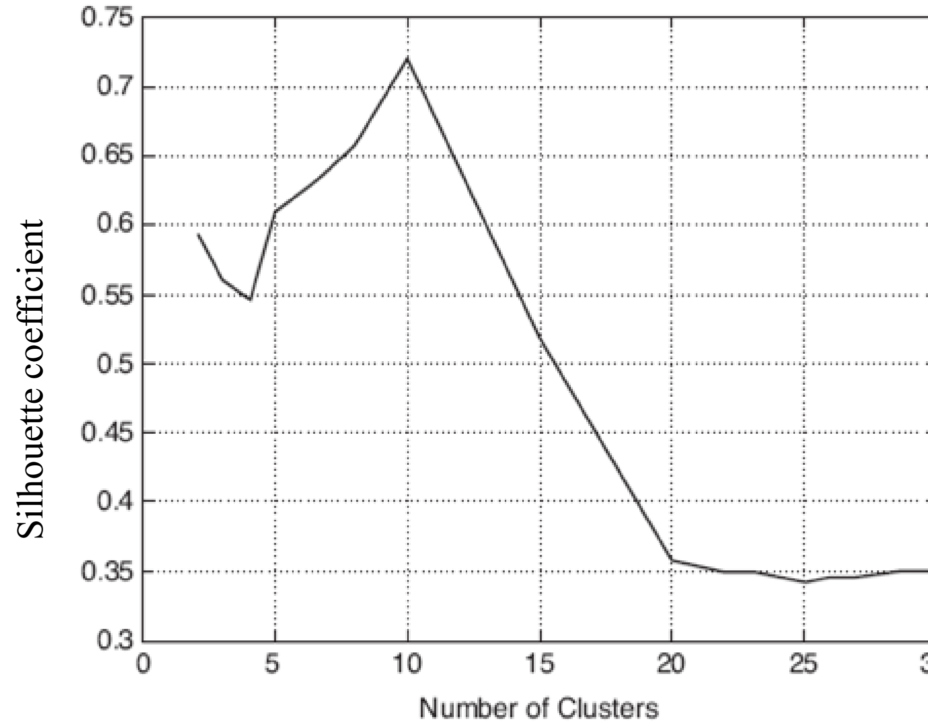


from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

7.2 Evaluation of clustering-algorithms

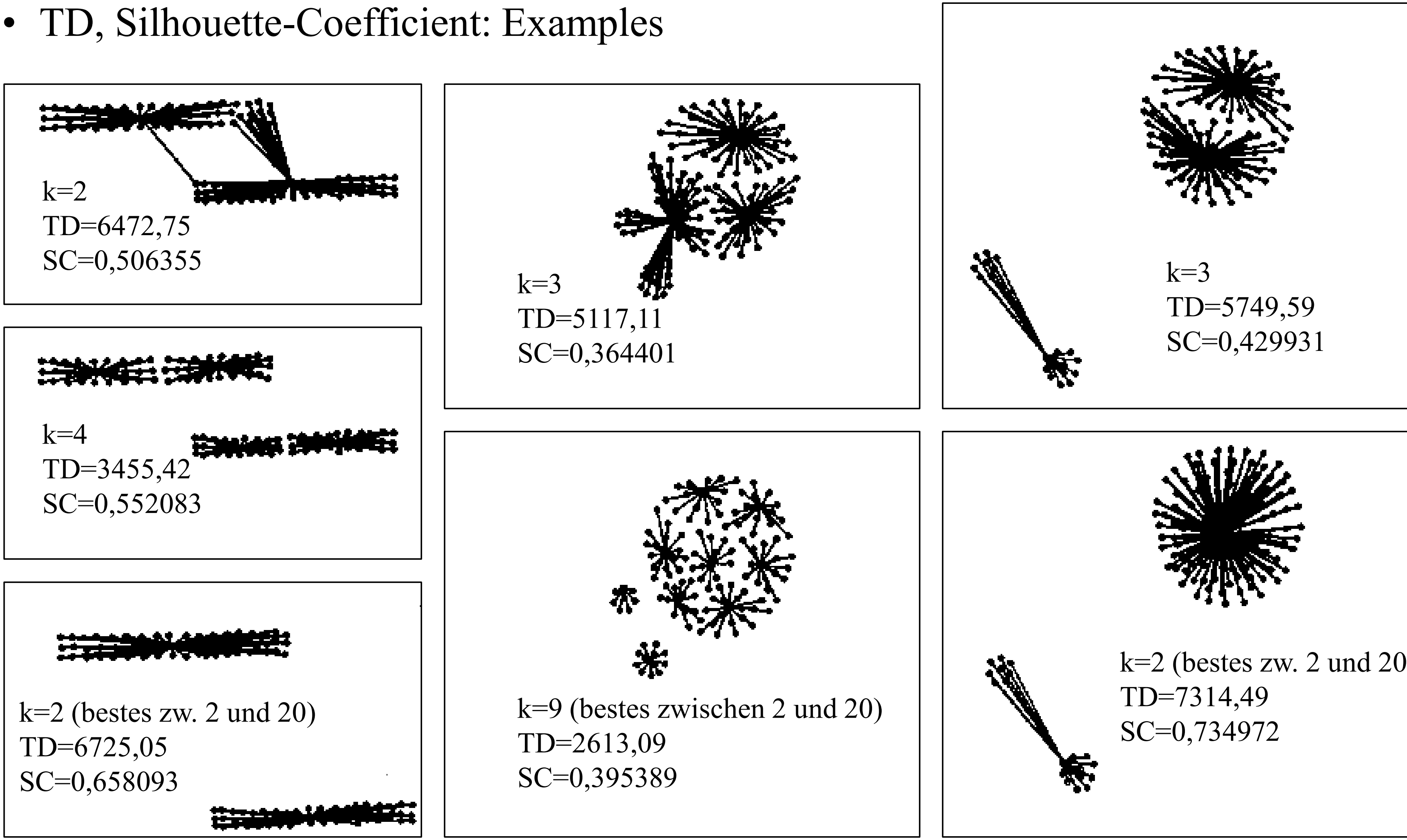
Comparison TD^2 – avg. silhouette-Coefficient for this data set





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

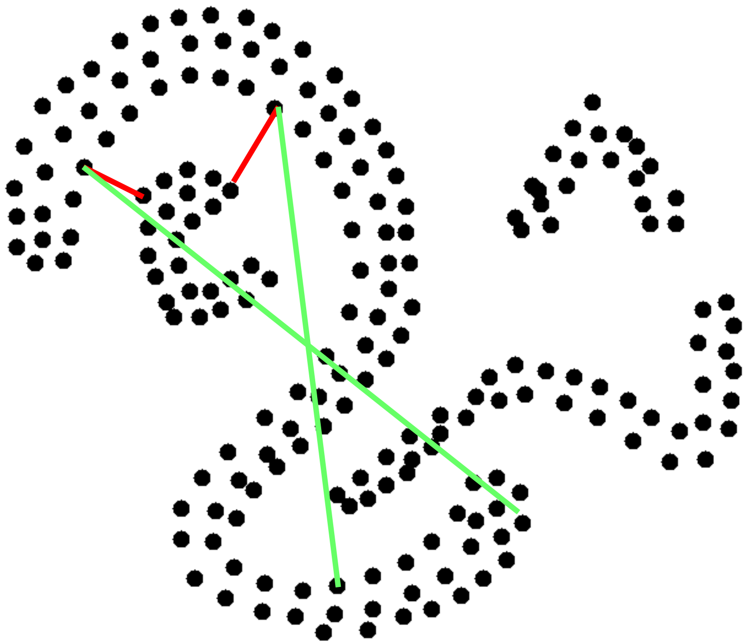
- TD, Silhouette-Coefficient: Examples



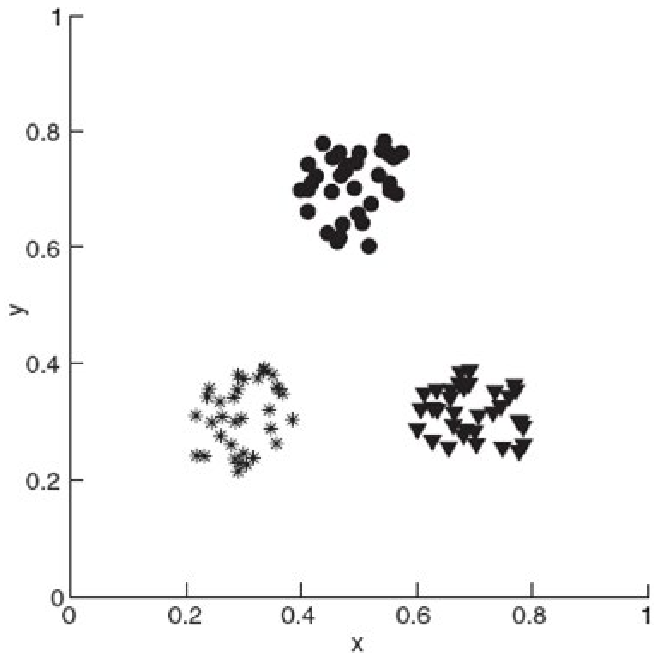
universität wien
Fakultät für Informatik

Cohesion and Separation

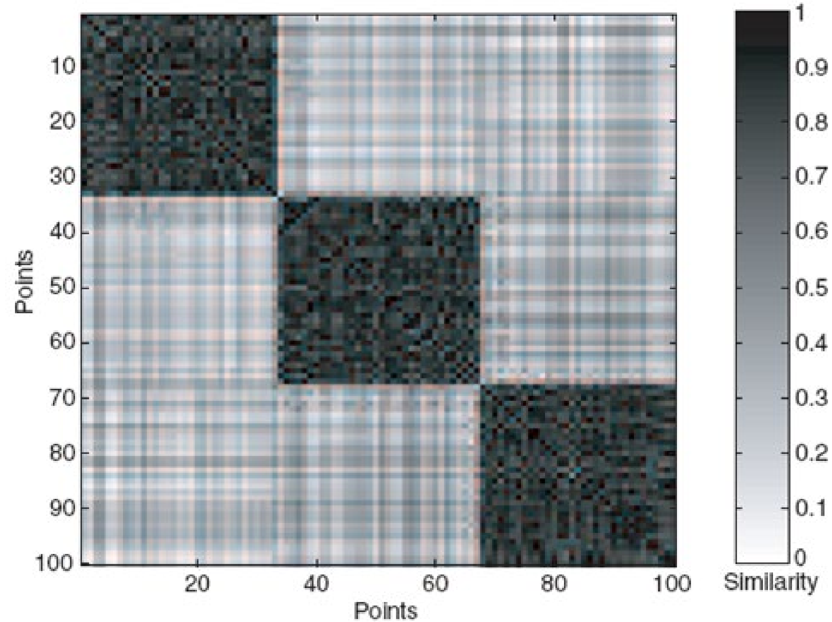
- Useful for spherical cluster, but not for cluster with “weird” shape



Evaluation of similarity matrices



Data set with well separated cluster

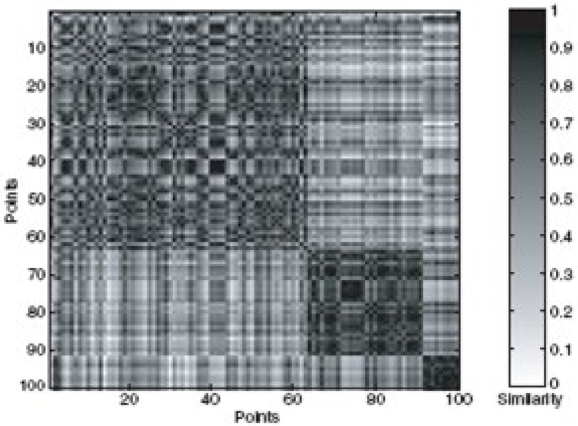


Similarity matrix (sortes with k-means labels)

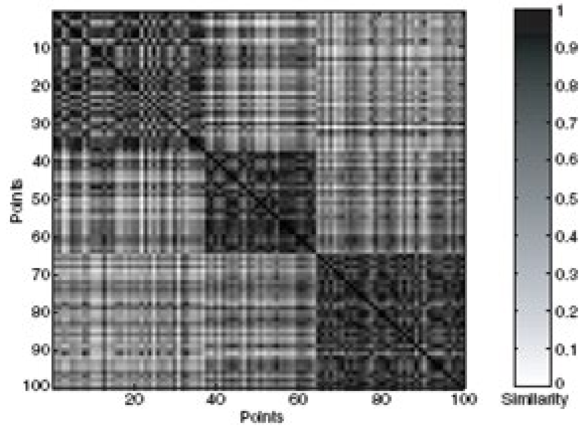
from: Tan, Steinbach, Kumar: Introduction to Data Mining (Pearson, 2006)

Evaluation of clustering-algorithms

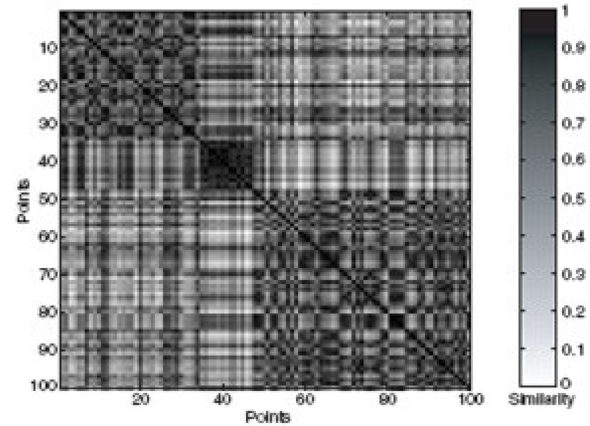
Similarity matrix:



DBSCAN



k-means



complete link

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

Evaluation of clustering-algorithms

Internal Evaluation - Problems:

- Adequacy of the cluster procedure for the given data
- Determinism?
- Finding k?
- How to compare different procedures against each other?
- Connection between the objective function of the cluster procedure and the evaluation function of the evaluation procedure

Evaluation of clustering-algorithms

External evaluation:

- Mapping between clustering result and given clusters (=classes) ("ground truth" or "gold standard")
  - why not just confusion matrix?
  - two basic approaches:
    - Mapping of sets of objects
    - Comparison of object pairs ("pair counting")
- Evaluation of the correspondence between given and found subsets
  - For mapping of sets of objects: information-theoretical measures
  - for comparison of object pairs: many dimensions, e.g. also classification-typical dimensions (F-measure etc.)

Mapping of object-sets:

- N objects
- Clustering U with R clusters $U_1, \ldots, U_R$
- Clustering V with C clusters $V_1, \ldots, V_C$
- n_{ij}: Number of elements in $U_i \cap V_j$
- $C \times R$ *contingency table*:

Information loss?

$$\left[n_{ij}\right]_{j=1\ldots C}^{i=1\ldots R}$$

$U \setminus V$	V_1	V_2	$\ldots$	V_C	Sum
U_1	n_{11}	n_{12}	$\cdots$	n_{1C}	a_1
U_2	n_{21}	n_{22}	$\cdots$	n_{2C}	a_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
U_R	n_{R1}	n_{R2}	$\cdots$	n_{RC}	a_R
Sum	b_1	b_2	$\cdots$	b_C	$\sum_{ij} n_{ij} = N$

?

Mapping of sets (example)

- Data set D: {1,2,3,4,5,6}
- Clustering U: {1,2,3}, {4,5}, {6}
- Clustering V: {1,2,4}, {3,5,6}

$U \setminus V$	{1,2,4}	{3,5,6}	Sum
{1,2,3}	2	1	3
{4,5}	1	1	2
{6}	0	1	1
Sum	3	3	6

Pair-Counting:

- N objects
- Clustering U with R clusters $U_1, \ldots, U_R$
- Clustering V with C clusters $V_1, \ldots, V_C$
  - a: Number of pairs of objects in the same cluster in U and V
  - b: Number of pairs of objects in the same cluster in U but not in V
  - c: Number of pairs of objects in the same cluster in V but not in U
  - d: Number of pairs of objects in different clusters in U and V
- 2×2 *contingency table*:

U\V	Pairs in same cluster	Pair in diff. Cluster
Pairs in same cluster	a	b
Pair in diff. Cluster	c	d

- Loss of information? Disjoint clusters?

$$a + b + c + d = \frac{N(N-1)}{2}$$

Evaluation of clustering-algorithms

- Pair-Counting (example):
- Data set D: {1, 2, 3, 4, 5, 6}
- Clustering U: {1, 2, 3}, {4, 5}, {6}
- Clustering V: {1, 2, 4}, {3, 5, 6}
- Pairs in D: {(1,2), (1,3), (1,4), (1,5), (1,6), (2,3), (2,4), (2,5), (2,6), (3,4), (3,5), (3,6), (4,5), (4,6), (5,6)}
- Pairs in U: {(1,2), (1,3), (2,3), (4,5)}
- Pairs in V: {(1,2), (1,4), (2,4), (3,5), (3,6), (5,6)}
- $a = |\text{ Pairs } U \cap \text{ Pairs } V| = |\{(1,2)\}| = 1$
- $b = |\text{ Pairs } U \setminus \text{ Pairs } V| = |\{(1,3), (2,3), (4,5)\}| = 3$
- $c = |\text{ Pairs } V \setminus \text{ Pairs } U| = |\{(1,4), (2,4), (3,5), (3,6), (5,6)\}| = 5$
- $d = |\text{ Pairs } D \setminus (\text{Pairs } U \cup \text{ Pairs } V)| = |\{(1,5), (1,6), (2,5), (2,6), (3,4), (4,6)\}| = 6$

U\V	Pairs in V	not pair in V
Pairs in U	1	3
Not pairs in U	5	6

Evaluation of clustering-algorithms

Evalution of set-mapping

- Precision/Recall?
  - Direction of Mapping?
  - Coverage?
- Entropy
- Mutual Information
- Normalized Mutual Information
- …

$U \setminus V$	V_1	V_2	V_3
U_1	10	0	0
U_2	12	1	3
U_3	8	5	7
U_4	25	8	8
U_5	15	7	7
U_6	20	0	0

Evaluation of clustering-algorithms

Evaluation of a pair-counting-matrix

- Precision, Recall, F-measure: like for classification
- Rand-Index
- Adjusted Rand Index (Hubert&Arabie)
- Jaccard-Index
- …

Cluster-wise Precision and Recall

WEKA output for K-means on Iris data set setting K to the number of classes.

Clustered Instances

```
0       61 ( 41%)
1       50 ( 33%)
2       39 ( 26%)
```

Classes to Clusters:

```
  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica
```

Cluster 0 <-- Iris-versicolor
Cluster 1 <-- Iris-setosa
Cluster 2 <-- Iris-virginica

Incorrectly clustered instances : 17.0 11.3333 %

https://www.cs.waikato.ac.nz/ml/weka/

https://archive.ics.uci.edu/ml/datasets/Iris

Cluster-wise Precision and Recall

universität wien
Fakultät für Informatik

$$Precision\ (class\ j, cluster\ k) = \frac{\#\ instances\ of\ class\ j\ in\ cluster\ k}{\#instances\ in\ cluster\ k}$$

How class pure is the cluster? Notice, singleton clusters always have precision 1

$$Recall\ (class\ j, cluster\ k) = \frac{\#\ instances\ of\ class\ j\ in\ cluster\ k}{\#\ instances\ in\ class\ j}$$

How comprehensive does the cluster represent the class? Notice, one large cluster Containg all the data is best in terms of recall.

Therefore: F-measure defined as harmonic mean of precision and recall.

Cluster-wise Precision and Recall

Classes to Clusters:

```
  0  1  2  <-- assigned to cluster
  0 50  0 | Iris-setosa
 47  0  3 | Iris-versicolor
 14  0 36 | Iris-virginica
```

Cluster 0 <-- Iris-versicolor, precision: 47/61(77%), recall: 47/50 (94%), F: 0.85
Cluster 1 <-- Iris-setosa, precision: 50/50 (100%), recall: 50/50 (100%), F: 1.0
Cluster 2 <-- Iris-virginica, precision: 36/39 (92%), recall 36/50 (72%), F: 0.81

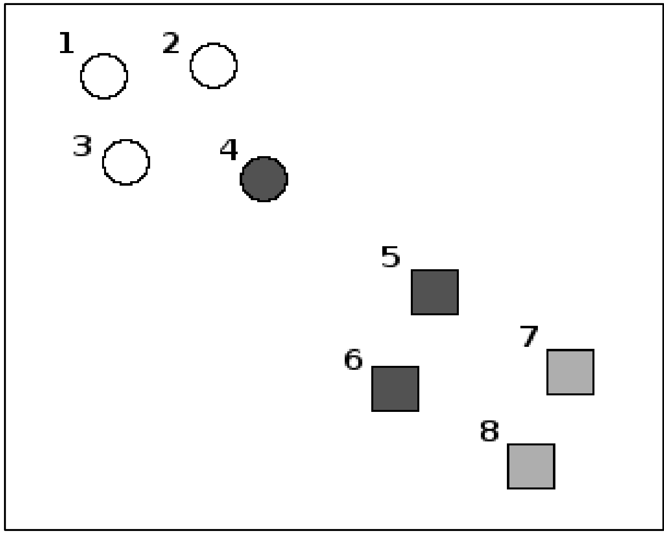
Need one number for the complete clustering:

- Average F-measure, here 0.88
- Useful: weighting with class size, here all classes have 50 objects, so not required.

Evaluation of clustering-algorithms

Rand Index:

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



a: #pairs in same class and same clusters

b: #pairs in same class, but different clusters

c: #pairs in different class, but same clusters

d: #pairs in different class and different clusters

Example:

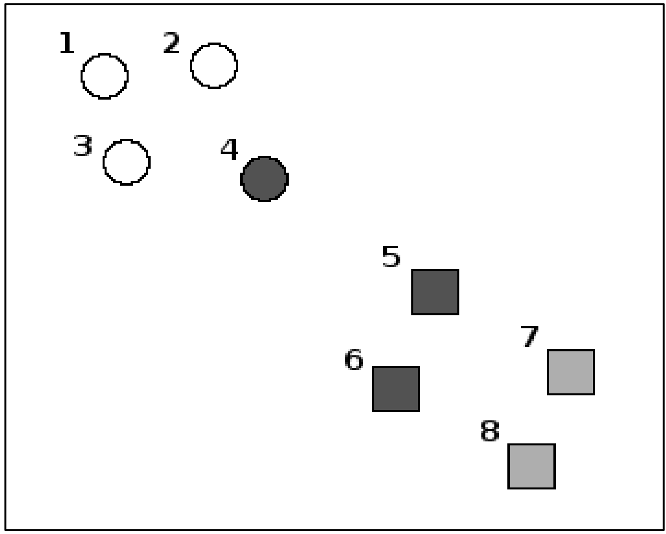
- 2 classes (Circle, Square)
- 3 cluster (Black, White, Gray)

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

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

Jaccard Coefficient

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



a: #pairs in same class and same clusters

b: #pairs in same class, but different clusters

c: #pairs in different class, but same clusters

d: #pairs in different class and different clusters

Example:

- 2 classes (Circle, Square)
- 3 cluster (Black, White, Gray)

a = 5; b = 7; c = 2

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

Adjusted Rand Index

Rand and Jaccard do not take into account the quality that can already be achieved by random solutions.

- Expected value is not 0, when comparing two random clusterings

- Adjustment for chance:

$$\text{Adjusted\_Criterion} = \frac{\text{Criterion} - E\left\{\text{Criterion}\right\}}{\text{Max\_Criterion} - E\left\{\text{Criterion}\right\}}$$

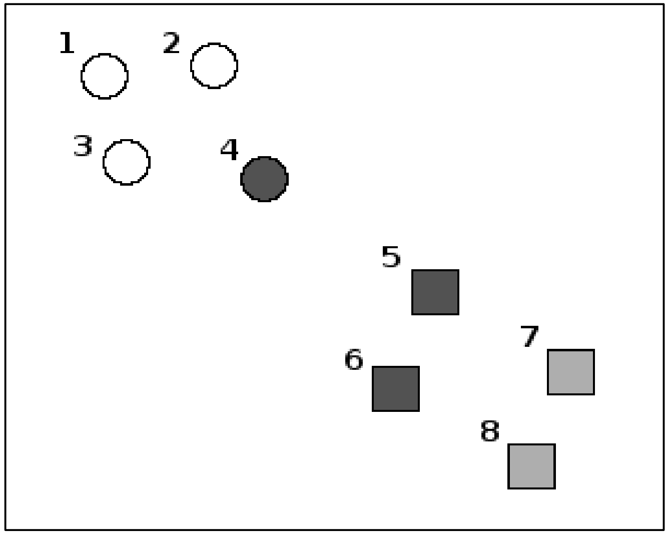
  - Hubert & Arabie (1985) analytically determined the expected value for the Rand Index and proposed the Adjusted Rand Index (ARI)

Evaluation of clustering-algorithms

- **Adjusted Rand Index**
- ARI can be written as:

$$ARI = \frac{a - \frac{(a+c)(a+b)}{M}}{\frac{(a+c)+(a+b)}{2} - \frac{(a+c)(a+b)}{M}}$$

with M = a + b + c + d



Example:

- 2 classes (Circle, Square)
- 3 cluster (Black, White, Gray)

a = 5; b = 7; c = 2; d = 14; M = 28

$$\mathbf{ARI} = \frac{5 - \frac{7*12}{28}}{\frac{7+12}{2} - \frac{7*12}{28}} = \mathbf{0.3793}$$

Normalized Mutual Information

$$NMI(classlabels, clusterIDs) = \frac{2\, I(classlabels, clusterIDs)}{H(classlabel) + H(clusterIDs)}$$

- Class labels and cluster Ids are categorical variables
- H represents the entropy

$$I(classlabels, clusterIDs) = \mathrm{H(classlabels)} - \mathrm{H(classlabels|clusterIDs)}$$

The mutual information tells us how much information we learn about the classlabels when we know the cluster IDs, does not favor small clusters.

Evaluation of clustering-algorithms

What does finding the classes mean?

- A data set can have several, different concept levels - which classes are found again? (Färber et al. 2010)
- Example: Amsterdam Library of Object Images (ALOI)
- 1000 objects
- same object = same class

7.2 Evaluation of clustering-algorithms

Feature for each object:

- different lighting angles
- different lighting colours
- different viewing angles (angle of rotation of the object)

Evaluation of clustering-algorithms

Possible concepts:

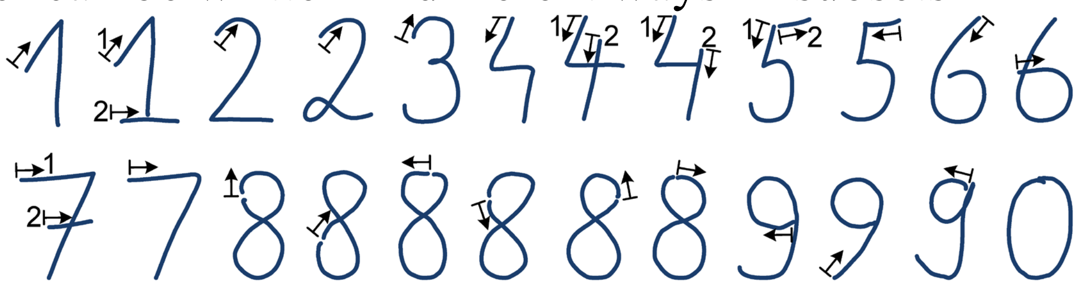
- same object in different colours (shape, object type)
- Different or similar objects in the same view
- dominant colour and/or shape
- …

Evaluation of clustering-algorithms

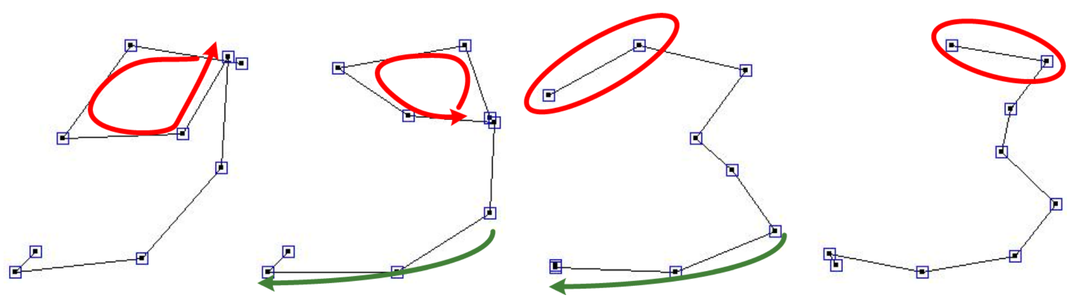
universität wien
Fakultät für Informatik

Example: **Pendigits data set**: handwritten digits

- Classes: Numbers 0-9
- Features: single (time) points in the course of the lettering
- meaningful concepts differ from the given classes:
  - a number can be written in different ways -> subsets



  - different digits can be rather similar



Evaluation of outlier-detection-algorithms

Evaluation of outlier procedures:

- Outlier vs. Inlier - a classification problem?
- Class Imbalance: The "Class" Outlier is much smaller, but often more important than the "Class" Inlier – difficult for:
  - Practice
  - Evalutation
- Many outlier methods do not provide a class decision, but rather outlier scores or factors – this enables a ranking of the objects.
- Usual evaluation scheme for ranked results: "Receiver Operating Characteristic“ (ROC)

Evaluation of outlier-detection-algorithms

Receiver Operating Characteristic (ROC)

$C(o) \setminus K(o)$	Outlier	Inlier
Outlier	TP	FN
Inlier	FP	TN

- Two-class confusion matrix:
  - "Outlier" is the class to discover
  - possible classification results
    - Outlier is recognized as Outlier: true positive (TP)
    - Outlier is classified as Inlier: false negative (FN)
    - Inlier is classified as Outlier: false positive (FP)
    - Inlier is recognized as Inlier: true negative (TN)
- Ranking: Continuum between strongest outlier and weakest outlier (= strongest inlier)
- Rated order: TP should come before FP if possible
- ROC: Graphical representation of TP rate vs. FP

Evaluation of outlier-detection-algorithms

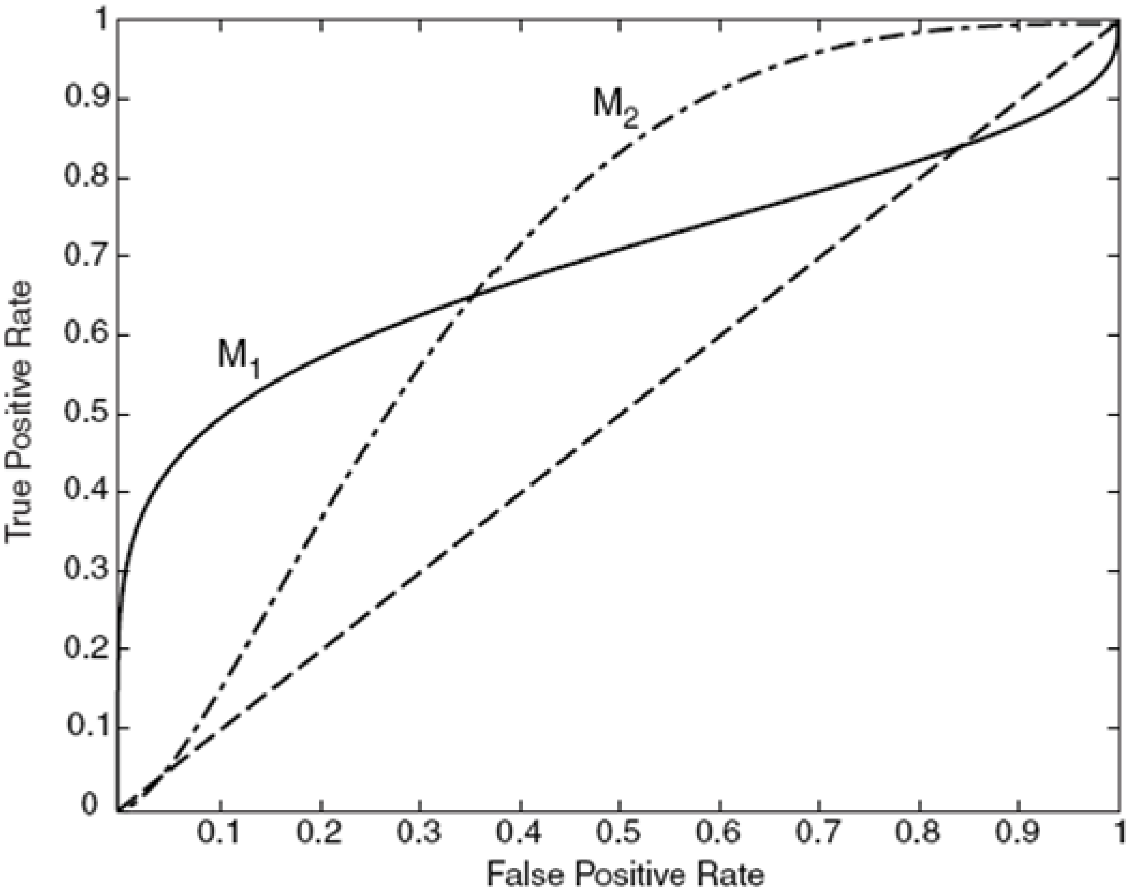
universität wien
Fakultät für Informatik

Receiver Operating Characteristic (ROC)

- each TP in the ranking: step upwards
- each FP in the ranking: step to the right
- random ranking?
- Comparison of two methods: Area under the ROC curve

  (ROC AUC)

  ($0 \leq$ ROC AUC ≤ 1)

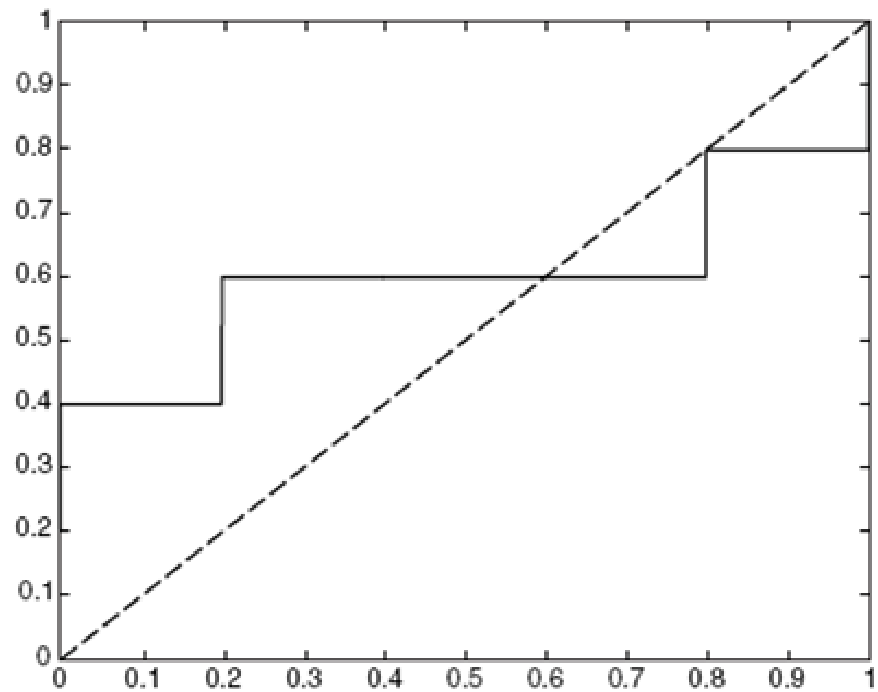


Evaluation of outlier-detection-algorithms

Example :

- 10 Objects
- Classes +/-
- Ranking:

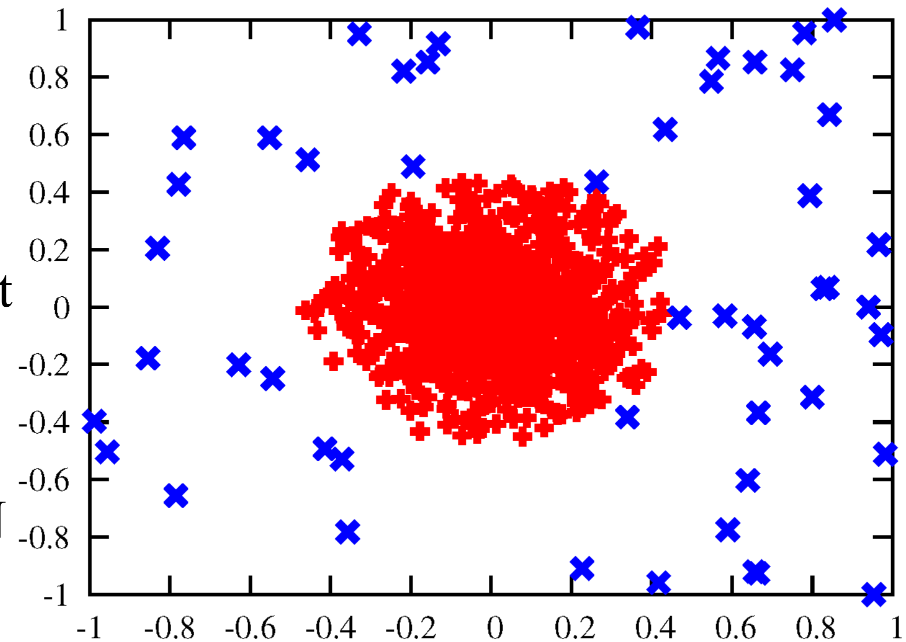
  Probability for class +

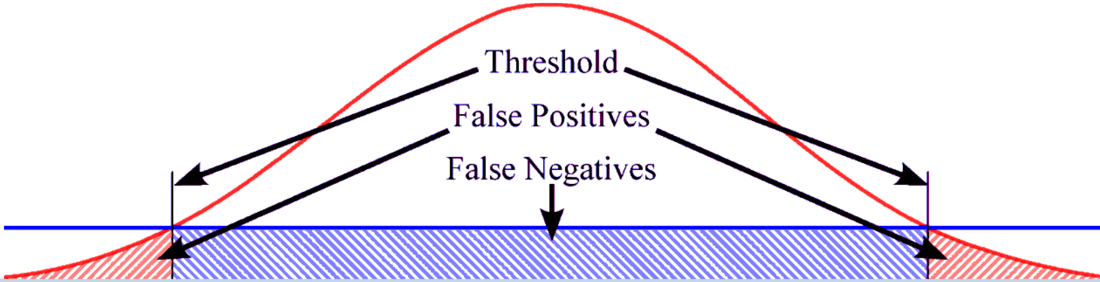


Class	+	−	+	−	−	−	+	−	+	+	
	0.25	0.43	0.53	0.76	0.85	0.85	0.85	0.87	0.93	0.95	1.00
TP	5	4	4	3	3	3	3	2	2	1	0
FP	5	5	4	4	3	2	1	1	0	0	0
TN	0	0	1	1	2	3	4	4	5	5	5
FN	0	1	1	2	2	2	2	3	3	4	5
TPR	1	0.8	0.8	0.6	0.6	0.6	0.6	0.4	0.4	0.2	0
FPR	1	1	0.8	0.8	0.6	0.4	0.2	0.2	0	0	0

Data for the evaluation of outlier methods

- How are the outliers, that are to be found by a procedure, defined in test data?
- synthetic data
  - Create a "normal" distribution (or multiple such) and a different lower density distribution
  - Even the most careful designed distribution will have FP and FN





Evaluation of outlier-detection-algorithms

Data for the evaluation of outlier methods

- How are the outliers, that are to be found by a procedure, defined in test data?

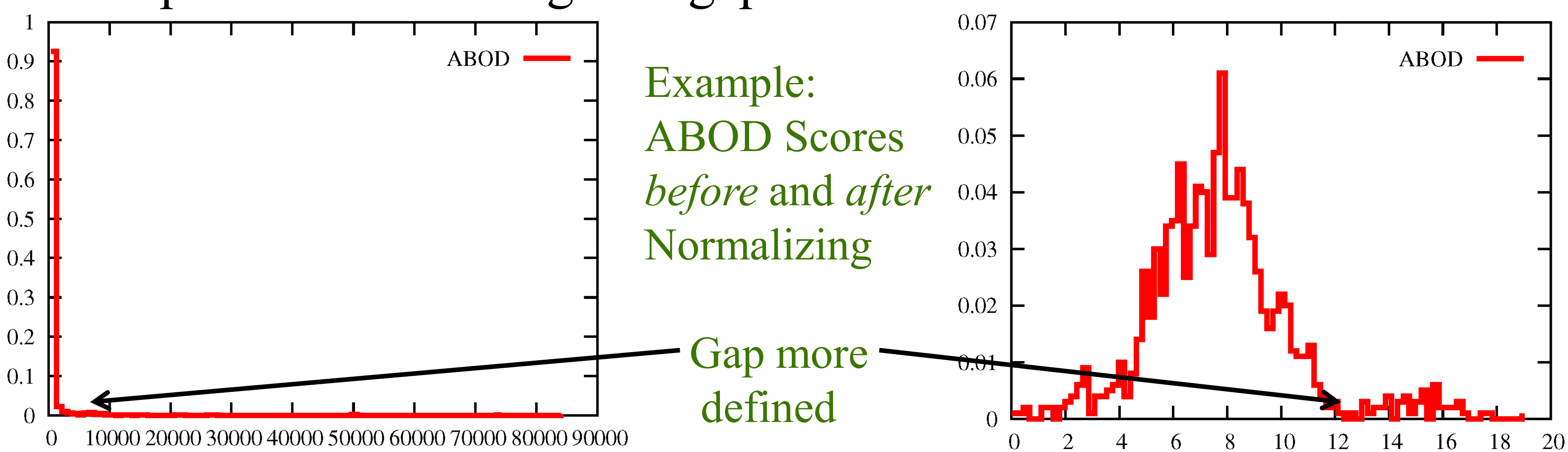
- real data
  - Hardly available
  - Classification problems, down-sampling of a class as outlier
  - the actual characteristics are unknown, a minimum of FP and FN cannot be ruled out

- relative performance comparison of different methods?
  - different methods find different outlier - correspondence to the characteristics of down-sampled class or borderline points of synthetic distributions often unclear (especially in high-dimensional data - no visual check possible!)

Evaluation of outlier-detection-algorithms

Evaluation of outlier procedures:

- Outlier scores must be interpreted meaningfully to make a decision (Outlier/Inlier)
- Approach: Mapping to "outlier probability" (Kriegel et al. 2011)
- Regularization/normalization of the outlier scores to 0-∞/0-1
- if possible increasing the "gap" between outlier and inlier



Evaluation of outlier-detection-algorithms

universität wien
Fakultät für Informatik

- possible evaluation including error costs (transferred from evaluation techniques for classification problems with imbalanced classes):
  - How bad is it to classify an outlier as an inlier (and vice versa)?
  - Cost weighting by probability of being outlier/inlier

$$\text{Cost} = \frac{1}{2}\sum_{x \in I} P(O \mid x) \cdot \frac{1}{|I|} + \frac{1}{2}\sum_{x \in O} P(I \mid x) \cdot \frac{1}{|O|}$$

What did you learn?

- Evaluation of results requires thorough analysis
- often no clear, absolute statement possible
- Comparison of results relative to each other (better/poorer): requires criterion
  - Comparison of results: Criterion must be appropriate to the problem
  - Comparison of procedures: Criterion must not systematically prefer individual procedures
- internal vs. external evaluation
- internal evaluation measures
  - Cohesion, separation: compactness, silhouette, similarity matrix
- external evaluation measures
  - mapping vs. pair counting
  - Precision, Recall, Edge Index, Jaccard, ARI
  - Outlier detection: Receiver Operating Characteristic
- Problem with the use of the "ground truth".