Pattern Recognition Lecture
Clustering Basics

Prof. Dr. Marcin Grzegorzek

Research Group for Pattern Recognition
Institute for Vision and Graphics
University of Siegen, Germany

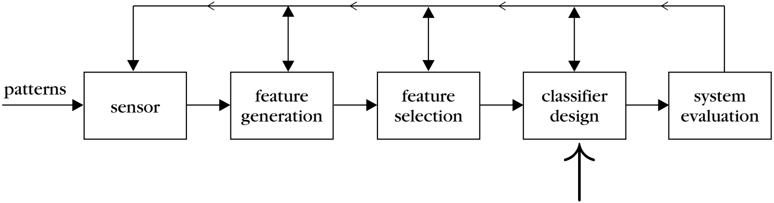
UNIVERSITÄT
SIEGEN

IVG

ZESS

Imaging
New Modalities
DFG-Graduiertenkolleg 1564

Pattern Recognition Chain



Overview

1 Introduction

2 Proximity Measures

Overview

1 Introduction

2 Proximity Measures

Unsupervised Learning in General

- Beginning from this lecture we will be dealing with the unsupervised case, i. e., the class labelling of the training patterns are not available

- Our major concern now is to organise the patterns into clusters (groups), which will allow us to discover similarities and differences among patterns and to derive useful conclusions about them.

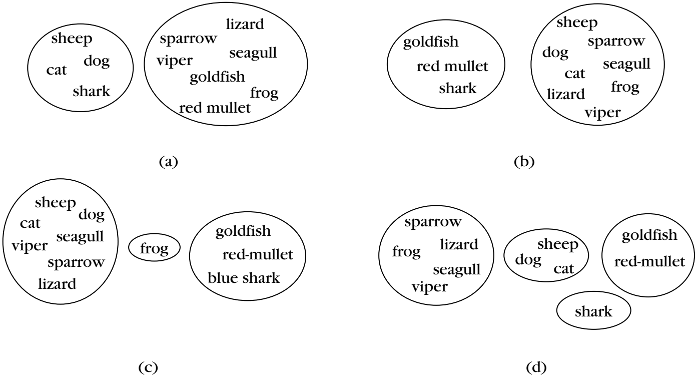
Different Names for Clustering

- Pattern Recognition: unsupervised learning or learning without a teacher (in our lecture it is called unsupervised algorithms for a certain reason...).
- Biology, Ecology: numerical taxonomy.
- Social Sciences: typology.
- Graph Theory: partition.

Introductory Example (1)

- Consider the following animals:
    - Mammals: sheep, dog, cat
    - Birds: sparrow, seagull
    - Reptiles: viper, lizard
    - Fish: goldfish, red mullet, blue shark
    - Amphibians: frog
- In order to organise them into clusters, we need to define a clustering criterion.

Introductory Example (2)



Resulting clusters if the clustering criterion is (a) the way the animals bear their progeny, (b) the existence of lungs, (c) the environment where the animals live, and (d) the way these animals bear their progeny and the existence of lungs.

Introductory Example (3)

- How many clusters would we get in the example, if the clustering criterion were the existence of a vertebral column?

- This example shows that the process of assigning objects to clusters may lead to very different results, depending on the specific criterion used for clustering.

Human View at Clustering

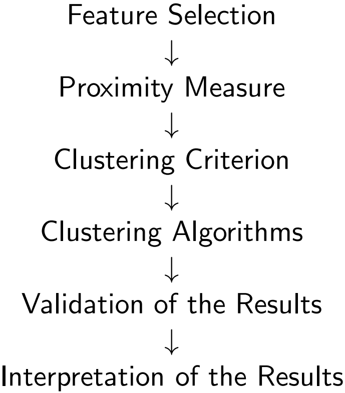
- Clustering is one of the most primitive mental activities of humans, used to handle the huge amount of information they receive every day.
- Processing every piece of information as a single entity would be impossible. Thus, humans tend to categorise entities (i. e., objects, persons, events) into clusters.
- For example, most humans “possess” a cluster “dog”.

Introduction

Proximity Measures

Basic Steps of Clustering Approaches

Assumption: patterns are represented by features, which form *l*-dimensional feature vectors.



Feature Selection

↓

Proximity Measure

↓

Clustering Criterion

↓

Clustering Algorithms

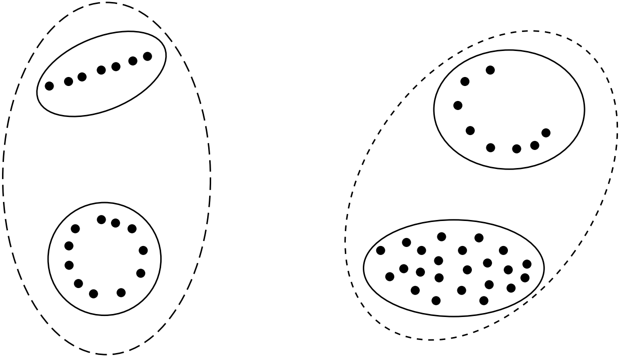
↓

Validation of the Results

↓

Interpretation of the Results

Subjectivity of Designing a Clustering Approach



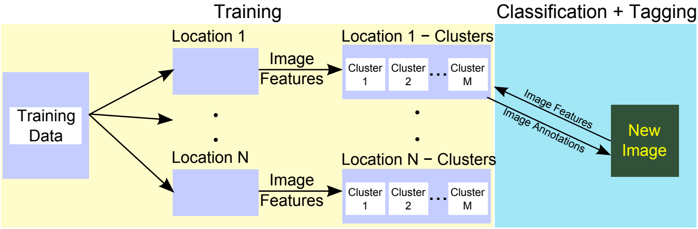
Depending on the sensibility of the clustering criterion, the clustering of the data above results in either two or four clusters.

Applications of Cluster Analysis

- Data Reduction
- Hypothesis Generation
- Hypothesis Testing
- Prediction Based on Groups

Tag Recommendation for Images

Framework for tag recommendation in social media. We used national capitals as locations and geographical coordinates, tags, and low-level features as image features.



R. Abbasi, M. Grzegorzek, and S. Staab. Large Scale Tag Recommendation Using Different Image Representations. In T.-S. Chua, Y. Kompatsiaris, B. Mrialdo, W. Haas, G. Thallinger, and W. Bailer, editors, *4th International Conference on Semantic and Digital Media Technologies*, pages 65–76, Graz, Austria, December 2009. Springer Berlin / Heidelberg.

Types of Features

- Nominal: possible values code states (sex of a person).
- Ordinal: arranged in meaningful order (grade in pattern recognition course: 1,2,3,4).
- Interval-Scaled: difference between two values is meaningful while their ratio is meaningless (Paris 10 and London 5 degrees Celsius).
- Ratio-Scaled: ratio between two values of a specific feature is meaningful (person 1 - 50 kg, person 2 - 100 kg).

Definitions of Clustering (1)

Clustering can be defined by the definition of the term cluster

Definition by Everitt et al. 2001

“Clusters are continuous regions of a feature space containing a relatively high density of points, separated from other high density regions by regions of relatively low density of points.”

Definitions of Clustering (2)

Formal Definition

Let X be our data set, that is $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. We define as an m-clustering of X, $\mathbb{R}$, the partition of X into m sets (clusters) $C_1, \ldots, C_m$, so that the following three conditions are met:

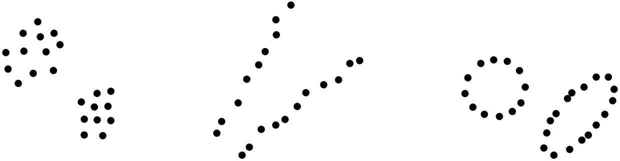
1. $C_i \neq \emptyset; \quad \forall i = 1, \ldots, m$
2. $\bigcup_{i=1}^{m} C_i = X$
3. $C_i \cap C_j = \emptyset; \quad \forall i \neq j; \quad i, j = 1, \ldots, m$

In addition, the vectors contained in a cluster C_i are "more similar" to each other and "less similar" to the feature vectors of the other clusters.

Definitions of Clustering (3)

- Quantifying the terms similar and dissimilar depends very much of the types of clusters involved.

- Other measures are required for compact clusters (a), others for elongated clusters (b), and different ones for shell-shaped clusters (c).



(a) (b) (c)

Overview

1 Introduction

2 Proximity Measures

Metric Dissimilarity Measure for Vector Pairs

Definition

A dissimilarity measure (DM) d for vector pairs from X is a function $d : X \times X \rightarrow \mathbb{R}$. It is called a **metric DM**, if it fulfils all of the following conditions:

1. $\exists d_0 \in \mathbb{R} : -\infty < d_0 \leq d(\mathbf{x}, \mathbf{y}) < +\infty; \quad \forall \mathbf{x}, \mathbf{y} \in X$
2. $d(\mathbf{x}, \mathbf{x}) = d_0; \quad \forall \mathbf{x} \in X$
3. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x}); \quad \forall \mathbf{x}, \mathbf{y} \in X$
4. $d(\mathbf{x}, \mathbf{y}) = d_0$ if and only if $\mathbf{x} = \mathbf{y}$
5. $d(\mathbf{x}, \mathbf{z}) \leq d(\mathbf{x}, \mathbf{y}) + d(\mathbf{y}, \mathbf{z}); \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in X$

Metric Similarity Measure for Vector Pairs

Definition

A similarity measure (SM) s for vector pairs from X is a function $s : X \times X \to \mathbb{R}$. It is called a **metric SM**, if it fulfils all of the following conditions:

1. $\exists s_0 \in \mathbb{R} : -\infty < s(\mathbf{x}, \mathbf{y}) \leq s_0 < +\infty; \quad \forall \mathbf{x}, \mathbf{y} \in X$
2. $s(\mathbf{x}, \mathbf{x}) = s_0; \quad \forall \mathbf{x} \in X$
3. $s(\mathbf{x}, \mathbf{y}) = s(\mathbf{y}, \mathbf{x}); \quad \forall \mathbf{x}, \mathbf{y} \in X$
4. $s(\mathbf{x}, \mathbf{y}) = s_0 \quad \text{if and only if} \quad \mathbf{x} = \mathbf{y}$
5. $s(\mathbf{x}, \mathbf{y})s(\mathbf{y}, \mathbf{z}) \leq [s(\mathbf{x}, \mathbf{y}) + s(\mathbf{y}, \mathbf{z})]s(\mathbf{x}, \mathbf{z}); \quad \forall \mathbf{x}, \mathbf{y}, \mathbf{z} \in X$

Proximity Measures between Sets of Vectors

- Sometimes the proximity measures are not computed for vector pairs but for pairs of sets of vectors from X.
- Let U be a set containing subsets of X. That is, $D_i \subset X$, $i = 1, \ldots, k$ and $U = \{D_1, \ldots, D_k\}$.
- A proximity measure δ on U is a function $\delta : U \times U \to \mathbb{R}$
- Conditions from the two previous slides can now be repeated with D_i and D_j in the place of $\mathbf{x}$ and $\mathbf{y}$ and U in the place of X in order to define the metric DM and the metric SM.

Quiz

We have two vectors $\mathbf{x} = [x_1, \ldots, x_l]^\mathsf{T}$ and $\mathbf{y} = [y_1, \ldots, y_l]^\mathsf{T}$ in an l-dimensional feature space.

A) Is the function $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{l} |x_i|$ a metric DM?

B) Is the function $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{l} |x_i - y_i|$ a metric DM?

C) Is the function $d(\mathbf{x}, \mathbf{y}) = \ln \left\{ \sum_{i=1}^{l} |x_i - y_i| \right\}$ a metric DM?

D) Is the function $d(\mathbf{x}, \mathbf{y}) = \sin \left\{ \sum_{i=1}^{l} |x_i - y_i| \right\}$ a metric DM?

E) Is the function $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{l} (x_i - y_i)$ a metric DM?

Proximity Measures between Two Points

- An example of a dissimilarity measure is here the weighted l_p metric DM, that is,

$$d_p(\mathbf{x}, \mathbf{y}) = \left(\sum_{i=1}^{l} \omega_i |x_i - y_i|^p \right)^{1/p}$$

- An example of a similarity measure is here the cosine similarity measure:

$$s_{\mathrm{cosine}}(\mathbf{x}, \mathbf{y}) = \frac{\mathbf{x}^{\mathsf{T}} \mathbf{y}}{\|\mathbf{x}\| \|\mathbf{y}\|}$$

Proximity Measures between a Point and a Set

- In many clustering schemes, a vector $\mathbf{x}$ is assigned to a cluster C taking into account the proximity between $\mathbf{x}$ and C, $\delta(\mathbf{x}, C)$. Some possibilities to compute this proximity measures:

$$\delta_{\max}^{ps}(\mathbf{x}, C) = \max_{\mathbf{y} \in C} \delta(\mathbf{x}, \mathbf{y})$$

$$\delta_{\min}^{ps}(\mathbf{x}, C) = \min_{\mathbf{y} \in C} \delta(\mathbf{x}, \mathbf{y})$$

$$\delta_{\mathrm{avg}}^{ps}(\mathbf{x}, C) = \frac{1}{n_C} \sum_{\mathbf{y} \in C} \delta(\mathbf{x}, \mathbf{y})$$

Introduction

Proximity Measures

Proximity Measures between Two Sets

- Most of the proximity measures δ^{ss} used for the comparison of sets are based on proximity measures between vectors. If D_i and D_j are two sets of vectors, the most common proximity functions are:

$$\delta^{ss}_{\max}(D_i, D_j) = \max_{\mathbf{x}\in D_i, \mathbf{y}\in D_j} \delta(\mathbf{x}, \mathbf{y})$$

$$\delta^{ss}_{\min}(D_i, D_j) = \min_{\mathbf{x}\in D_i, \mathbf{y}\in D_j} \delta(\mathbf{x}, \mathbf{y})$$

$$\delta^{ss}_{\mathrm{avg}}(D_i, D_j) = \frac{1}{n_{D_i} n_{D_j}} \sum_{\mathbf{x}\in D_i} \sum_{\mathbf{y}\in D_j} \delta(\mathbf{x}, \mathbf{y})$$