10. Tasks in Machine Learning

Contents:

Clustering

School of Electronic and Computer Engineering
Peking University

Wang Wenmin

10.3. Clustering

Contents:

What is Clustering 什么是聚类

- A longer description 较长描述

  Clustering is the task of grouping a set of objects in such a way that objects in the same group are more similar to each other than to those in other groups.

  聚类是以这样的一种方式将对象进行分组的任务，即同一组中的对象彼此之间比其他组中的对象更相似。

- A shorter description 较短描述

  The process of organizing objects into groups whose members are similar in some way.

  将对象进行分组的过程，组内成员具有某种方式的相似性。

- A very short description 极简描述

  To group data objects.

  将数据对象分组。

Clustering vs. Classification 聚类与分类

- Similarity 相似性
  Groups or classes
- Difference 差异性
  As shown in the following table 如下表所示

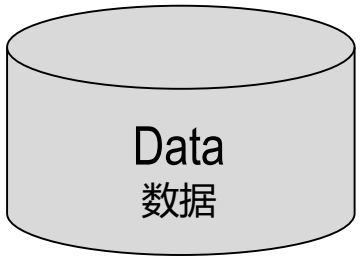
Clustering 聚类	Classification 分类
To identify similar groups for input objects 给输入对象**标识相似的组。**	To assign pre-defined classes for input items 给输入项**分派预定义的类。**
Without training data. **没有训练数据。**	With training data. **有训练数据。**
Clusters are discovered based on distances, density, etc. 基于距离、密度等发现类聚。	Classifiers need to have a high accuracy for classification. 分类器需要具有较高的分类精度。

Grouping Input Data into Same Cluster 将输入数据分成相同的类聚

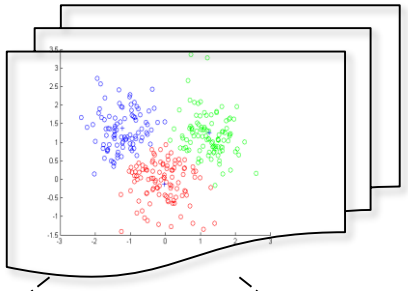
Clustering Algorithm
聚类算法

Input: No labeled.
输入：无标注的

Output: Grouped
输出：分组的



Cluster Analysis
聚类分析



Clusters
类聚

	Unseen
1	X
2	X
3	X
4	X

...

Unseen Clusters
未知类聚

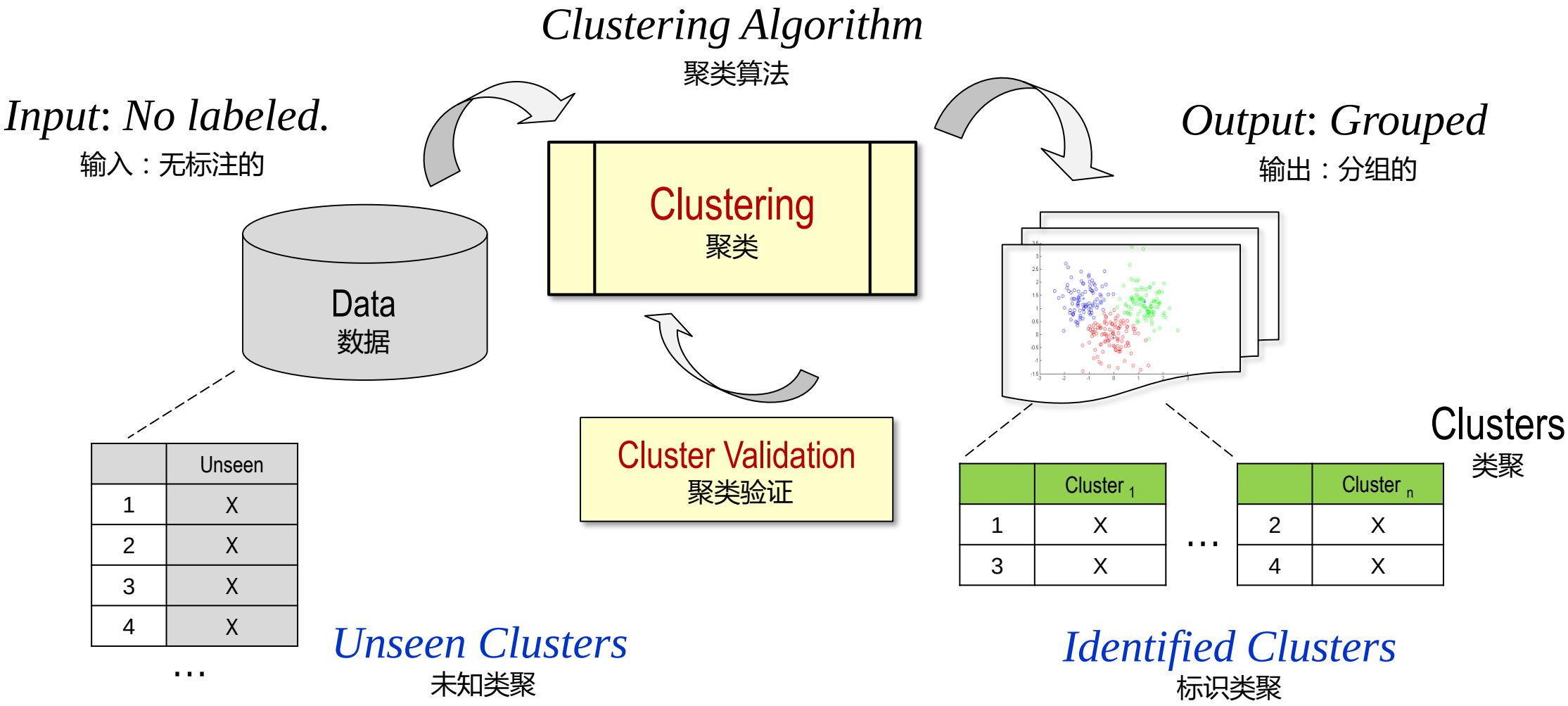
	Cluster $_1$
1	X
3	X

...

	Cluster $_n$
2	X
4	X

Identified Clusters
标识类聚

Two Key Steps in Clustering Procedure 聚类过程中的两个重要步骤



A Formal Description of Clustering 一种聚类的形式化描述

Let $\mathbb{R}^n$ $(n \geqq 1)$ denote a set of n-dimensional real-valued vectors, input space $\mathcal{X}$ is a subset of $\mathbb{R}^n$, output space $\mathcal{Y}$ is a set of unknown clusters, D is an unknown distribution over $\mathcal{X} \times \mathcal{Y}$, then:

设$\mathbb{R}^n$ $(n \geqq 1)$ 表示一个n维实数向量集，输入空间$\mathcal{X}$是$\mathbb{R}^n$的子集，输出空间$\mathcal{Y}$是一组未知的类聚， D是$\mathcal{X} \times \mathcal{Y}$笛卡尔乘积上的未知分布，则：

- Let a clustering function: 设聚类函数

$$h : \mathcal{X} \to \mathcal{Y} \text{ and } h \in H$$

- Clustering: 聚类

  Given a testing set of unknown clusters:

  给定一个未知类聚的测试集：

$$\mathcal{X} = \{x^{(i)} \mid x \in \mathcal{Y}, i \in [1, m]\}$$

  Using the clustering function determined at above to analyze the clustering results:

  采用上述确定的聚类函数来分析聚类结果：

$$\mathcal{Y} = h(\mathcal{X}) = \{y^{(i)} \mid y \in \mathcal{Y}, i \in [1, n], h(x) = y\}$$

10.3. Clustering

Contents:

Typical Approaches of Clustering Algorithm 聚类算法的典型方法

- 1) Connectivity-based clustering 基于连接性聚类

  Also known as hierarchical clustering, based on the distance between objects.

  也被称为基于对象间距离的层次聚类。

- 2) Centroid-based clustering 基于中心点聚类

  To find the k cluster centers and assign the objects to nearest cluster center.

  发现k个类聚中心并将对象分配到最近的类聚中心点。

- 3) Distribution-based clustering 基于分布聚类

  Clusters can be defined as objects belonging most likely to the same distribution.

  类聚可被定义为恰好属于同一分布的对象群。

- 4) Density-based clustering 基于密度聚类

  To group objects into one cluster if they are connected by densely populated area.

  将稠密区域连接的对象组成一个类聚。

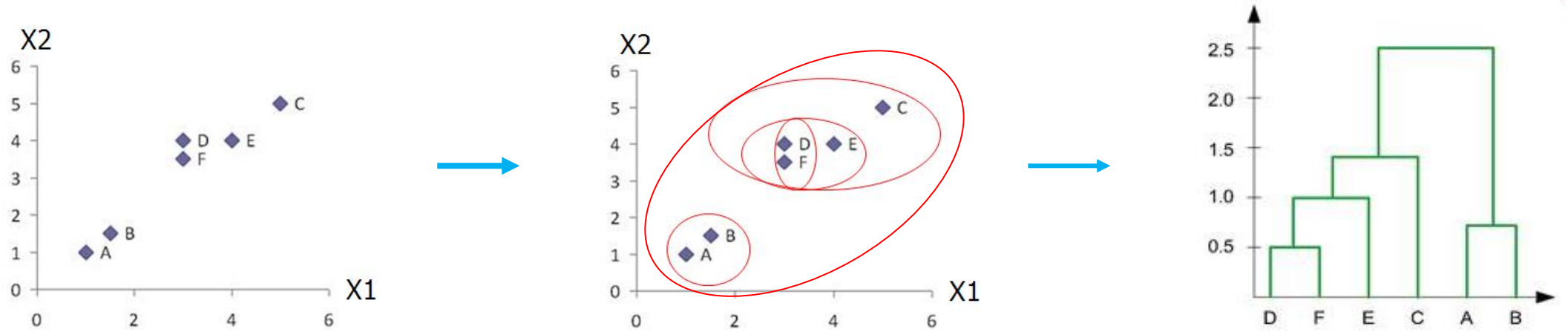
1) Connectivity-based clustering 基于连接性聚类

- Based on the core idea of objects being more related to nearby objects than to objects farther away.

  基于这样一个核心理念：对象与其附近的对象更相关，而不是较远的对象。

- Creating a hierarchical decomposition of the set of data objects using some criterion.

  采用某种准则来创建数据对象集的层次分解。



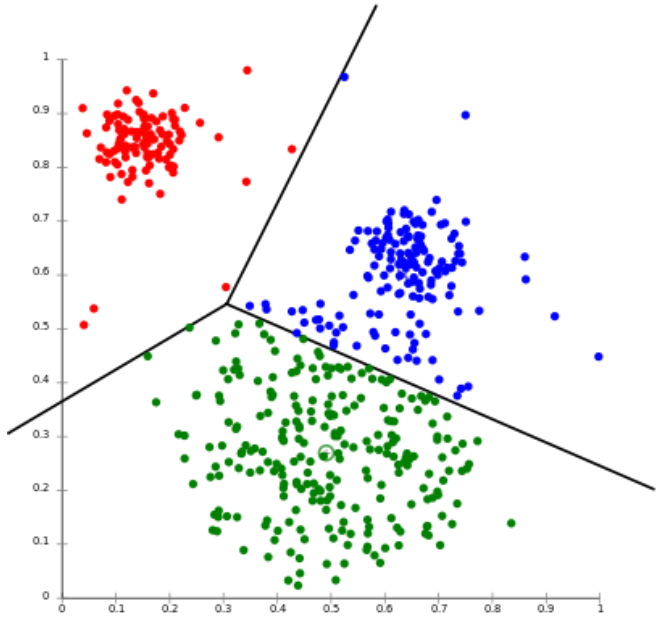
Typical algorithms: AGNES (Agglomerative NESting), DIANA (Divisive Analysis), ……

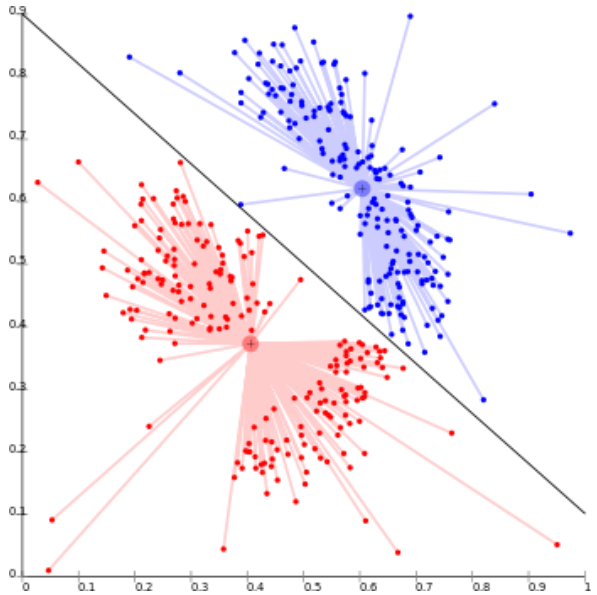
典型算法：AGNES (集聚嵌套), DIANA (分裂分析), ...

2) Centroid-based clustering 基于中心点聚类

- Constructing various partitions and then evaluating them by some criterion, e.g., minimizing the sum of square distance cost.

  构建各种不同的分区，再根据某种准则（例如最小平方距离代价之和）对其进行评价。





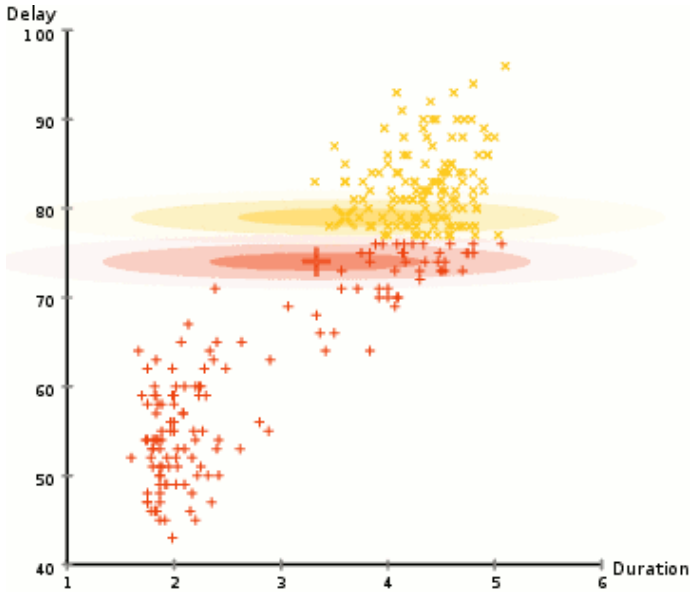
Typical algorithms: k-means, k-medoids, ……

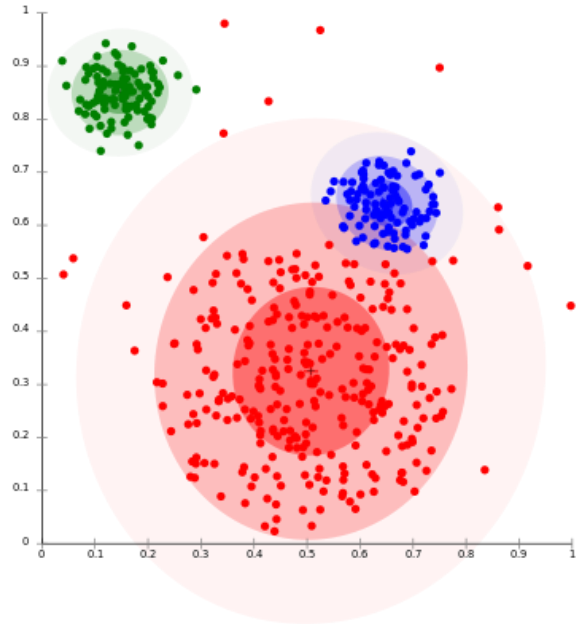
典型算法：k-均值, k-中心点, ...

3) Distribution-based clustering 基于分布聚类

- Clusters are modeled using statistical distributions, such as multivariate normal distributions.

  采用统计分布（诸如多元正态分布）对类聚进行建模。





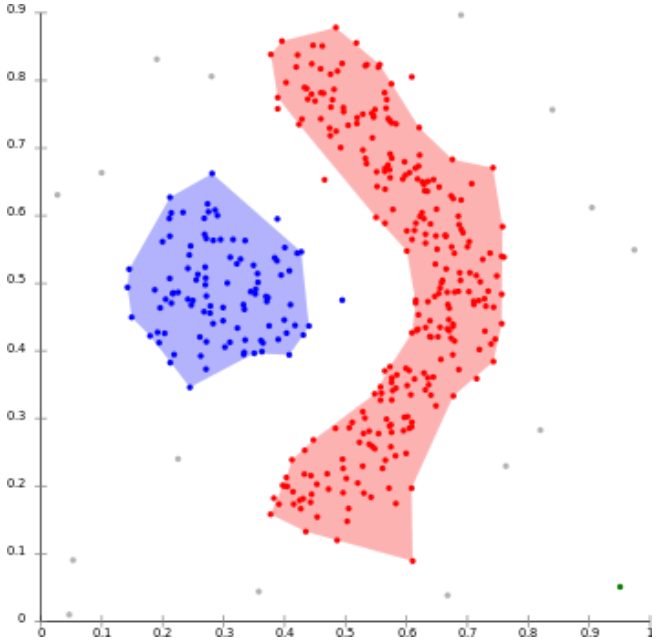
Typical algorithms: Expectation-maximization, ……

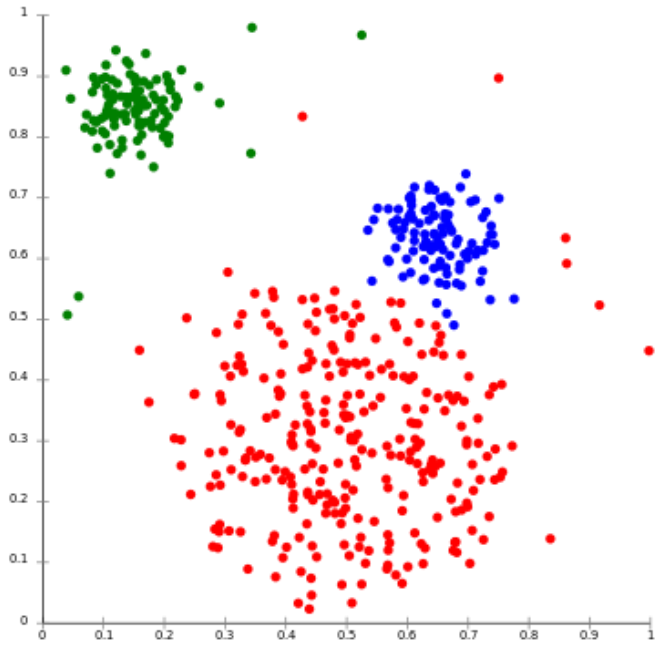
典型算法：期望最大化, ...

4) Density-based clustering 基于密度聚类

- Clusters are defined as areas of higher density than the remainder of the data set.

  类聚被定义为比数据集其余部分密度更高的区域。





Typical algorithms: DBSCAN (Density-Based Spatial Clustering of Applications with Noise), ……

典型算法：DBSCAN (基于密度的噪声应用空间聚类), ...

Case Study: Clustering by density peaks 根据密度峰值聚类

- Cluster centers are characterized by
  - 1) a higher density than their neighbors,
  - 2) a larger distance from points with higher densities.

  类聚中心点的特性是：1) 密度高于其相邻点，2) 距离大于其它较高密度点。

Source: "Clustering by fast search and find of density peaks", *SCIENCE, Vol. 344, Jun. 27 2014.*

- The features of the clustering method are:

  该聚类方法的特点：
  - the number of clusters arises intuitively,

    直观地得到类聚的个数，
  - outliers are automatically spotted and excluded,

    自动地发现和排除离群点，
  - clusters are recognized regardless of their shape, and space dimensionality.

    无论其形状以及空间的维度，类聚都能被识别。

Case Study: Clustering by density peaks 根据密度峰值聚类

Local density:
局部密度:

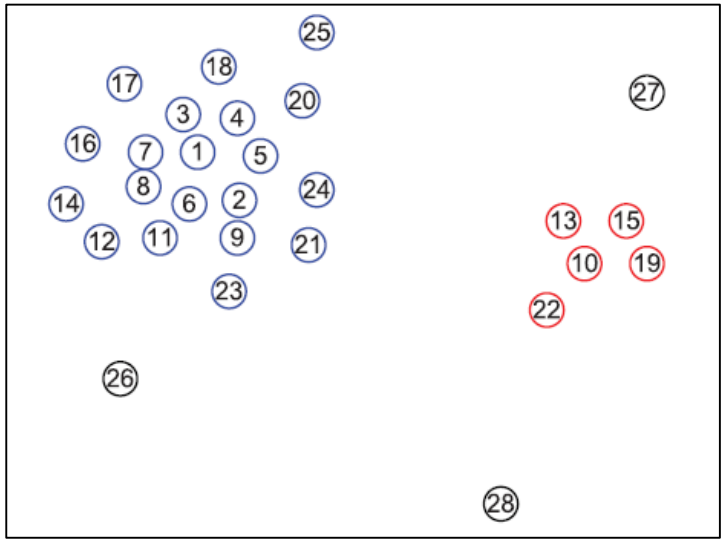
$$\rho_i = \sum_j \chi(d_{ij} - d_c) \qquad \chi(x) = \begin{cases} 1 & \text{if } x < 0 \\ 0 & \text{otherwise} \end{cases}$$

Minimum distance:
最小距离:

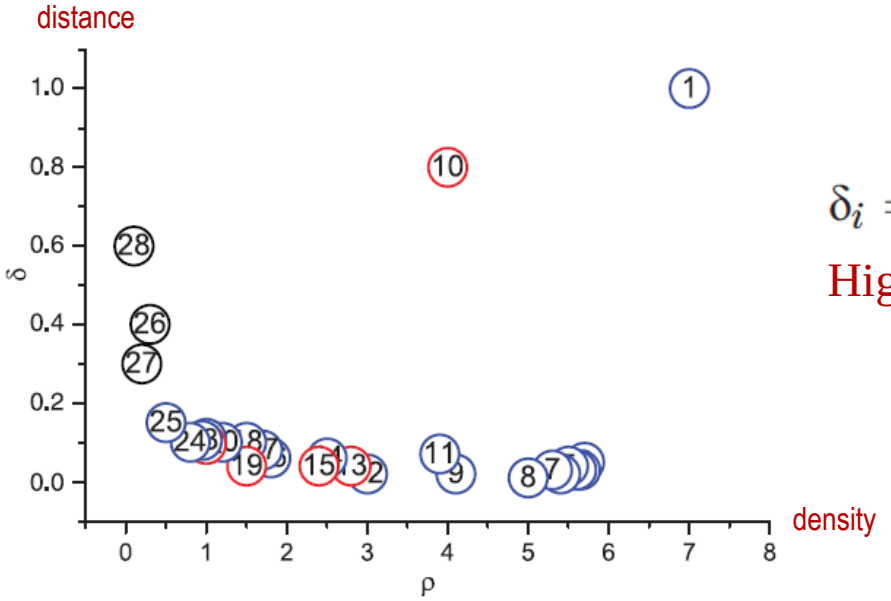
$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij})$$

where, d_{ij} : the distances between data points 数据点之间的距离

d_c : cutoff distance. 截断距离



Data (28 points) in decreasing density.
密度降排表示的数据（28个点）



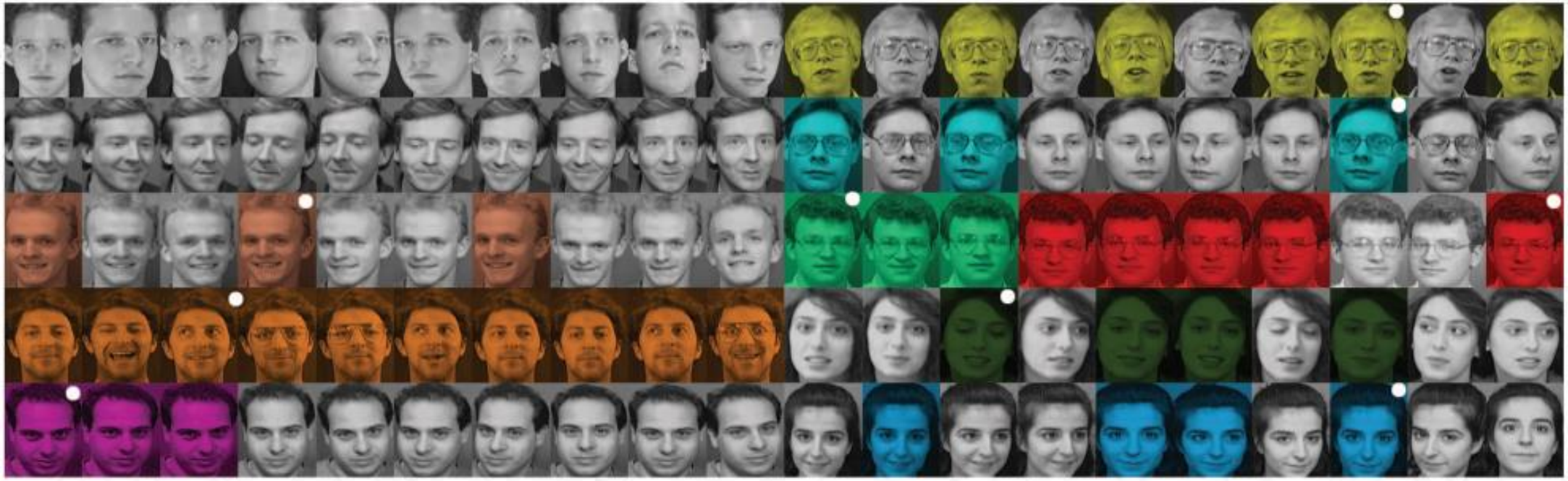
$$\delta_i = \max_j(d_{ij})$$

Highest density
最高密度

Decision graph calculated local density and distance
计算局部密度和距离后的决策图

Case Study: Clustering by density peaks 根据密度峰值聚类

- Clustering analysis of the Olivetti Face Database. 人脸数据库Olivetti的聚类分析



Pictorial representation of the cluster assignations for the first 100 images. Faces with the same color belong to the same cluster, whereas gray images are not assigned to any cluster. Cluster centers are labeled with white circles.

前100幅图像类聚分配的图片表示。具有同样颜色的人脸属于同一个类聚，而灰色图像表示没被分配到任何类聚。类聚中心标有白色圆圈。

10.3. Clustering

Contents:

Typical Applications of Clustering 聚类的典型应用

- Medicine 医学
  - Medical imaging 医学影像
- Business and marketing 商务和营销
  - Grouping of customers 顾客分组
  - Grouping of shopping items 购物商品分组
- World wide web 万维网
  - Social network analysis 社交网络分析
  - Search result grouping 搜索结果分组
- Computer science 计算机科学
  - Image segmentation 图像分割
  - Recommender systems 推荐系统

Typical Algorithms of Clustering 典型的聚类算法

- k-means
- k-modes
- PAM
- CLARA
- FCM
- BIRCH
- CURE
- ROCK
- Chameleon
- Echidna
- DBSCAN
- DBCLASD
- OPTICS
- DENCLUE
- Wave-Cluster
- CLIQUE
- STING
- OptiGrid
- EM
- CLASSIT
- COBWEB
- SOMs

Thank you for your attention!

AI