Technology
Science
Information
Networks
Computing

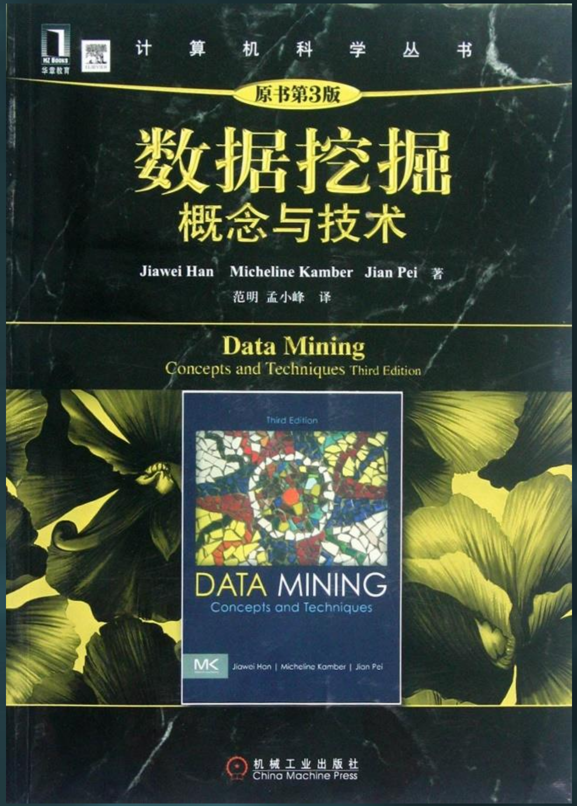
TSINC

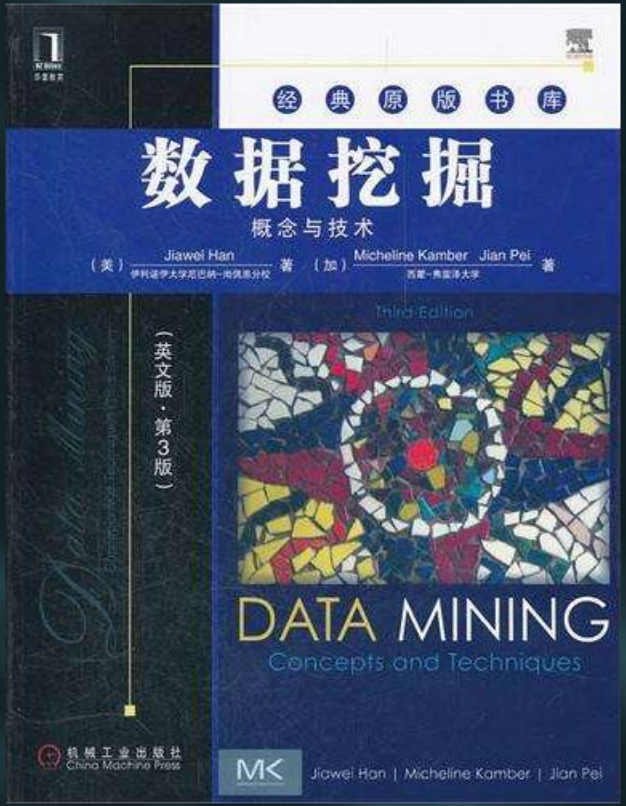
Lecturer: Ting Wang (王挺)

利物浦大学计算机博士
清华大学计算机博士后
电子信息技术高级工程师
上海外国语大学网络与新媒体副教授
浙江清华长三角研究院海纳认知与智能研究中心主任

Chapter 10

Clustering





Chapter 10 Clustering

1. What is Clustering

- A **cluster is** a collection of data objects that are *similar* to one another within the same cluster and are *dissimilar* to the objects in other clusters.
- The process of grouping a set of physical or abstract objects into classes of *similar* objects is called **clustering**
- **Unsupervised learning:** no predefined classes

Chapter 10 Clustering

2. The quality of a clustering method depends on

- the similarity measure used by the method
- its implementation,
- Its ability to discover some or all of the hidden patterns

Similarity measure methods:

- distance-based methods can often take advantage of optimization techniques
- density- and continuity-based methods can often find clusters of arbitrary shape

Chapter 10 Clustering

3. Compare clustering methods

- Partitioning criteria
  - Single level vs. hierarchical partitioning (often, multi-level hierarchical partitioning is desirable)
- Separation of clusters
  - Exclusive (e.g., one customer belongs to only one region) vs. non-exclusive (e.g., one document may belong to more than one class)
- Similarity measure
  - Distance-based (e.g., Euclidian, road network, vector) vs. connectivity-based (e.g., density or contiguity)
- Clustering space
  - Full space (often when low dimensional) vs. subspaces (often in high-dimensional clustering)

Chapter 10 Clustering

4. Partitioning

first creates an initial set of *k* partitions, where parameter*k* is the number of partitions to construct. It then uses an *iterative relocation technique* that attempts to improve the partitioning by moving objects from one group to another. Typical partitioning methods include *k*-means, *k*-medoids, and CLARANS.

Chapter 10 Clustering

(1)K-Mean

Algorithm: ***k*-means.** The *k*-means algorithm for partitioning, where each cluster's center is represented by the mean value of the objects in the cluster.

Input:

- *k*: the number of clusters,
- *D*: a data set containing *n* objects.

Output: A set of *k* clusters.

Method:

(1) arbitrarily choose *k* objects from *D* as the initial cluster centers;
(2) **repeat**
(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
(4) update the cluster means, that is, calculate the mean value of the objects for each cluster;
(5) **until** no change;

The *k*-means partitioning algorithm.

Chapter 10 Clustering

(2)K-Medoids, PAM

Algorithm: ***k*-medoids.** PAM, a k-medoids algorithm for partitioning based on medoid or central objects.

Input:

- k: the number of clusters,
- D: a data set containing n objects.

Output: A set of k clusters.

Method:

(1) arbitrarily choose k objects in D as the initial representative objects or seeds;
(2) **repeat**
(3) assign each remaining object to the cluster with the nearest representative object;
(4) randomly select a nonrepresentative object, o_{random};
(5) compute the total cost, S, of swapping representative object, o_j, with o_{random};
(6) **if** $S < 0$ **then** swap o_j with o_{random} to form the new set of k representative objects;
(7) **until** no change;

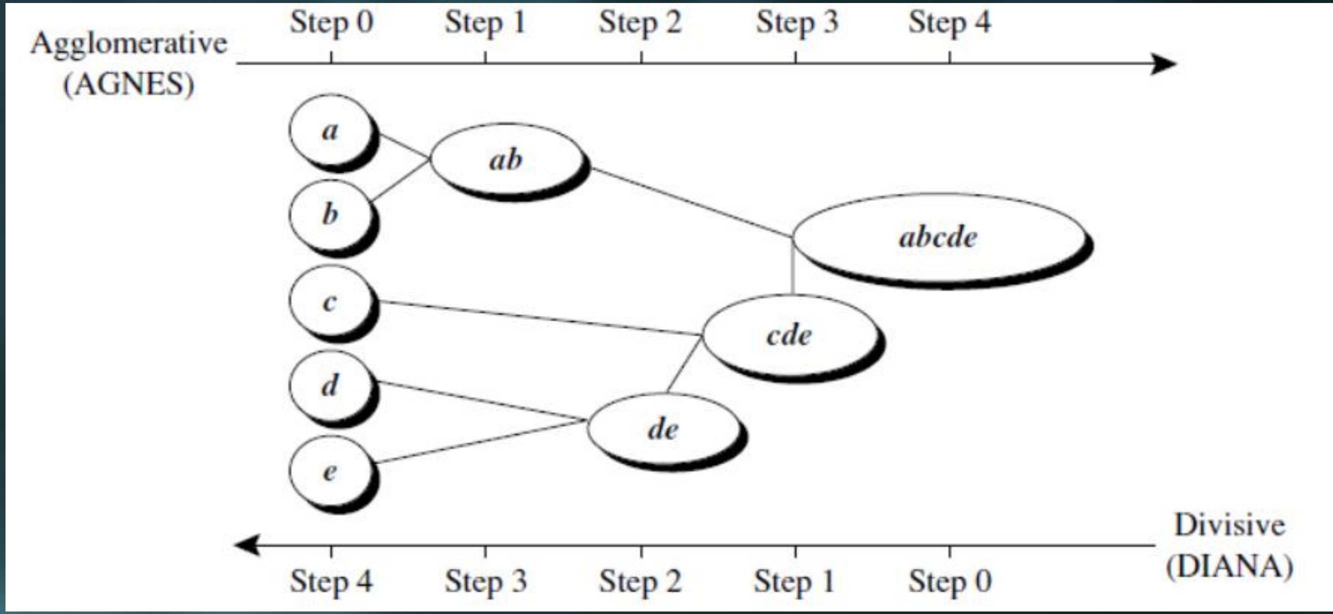
PAM, a k-medoids partitioning algorithm.

Chapter 10 Clustering

5. Hierarchical

creates a hierarchical decomposition of the given set of data objects. The method can be classified as being either *agglomerative* (*bottom-up*) or *divisive* (*top-down*), based on how the hierarchical decomposition is formed. To compensate for the rigidity of *merge* or *split*, the quality of hierarchical agglomeration can be improved by analyzing object linkages at each hierarchical partitioning (e.g., in Chameleon), or by first performing *microclustering* (that is, grouping objects into "microclusters") and then operating on the microclusters with other clustering techniques such as iterative relocation (as in BIRCH).

AGNES and DIANA



Chapter 10 Clustering

6. Distance Measures in clusters

Minimum distance: $dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} \{|p - p'|\}$

Maximum distance: $dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} \{|p - p'|\}$

Mean distance: $dist_{mean}(C_i, C_j) = |m_i - m_j|$

Average distance: $dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'|$

Chapter 10 Clustering

7. Density-Based

clusters objects based on the notion of density. It grows clusters either according to the density of neighborhood objects (e.g., in DBSCAN) or according to a density function (e.g., in DENCLUE). OPTICS is a density-based method that generates an augmented ordering of the data’s clustering structure.

8. DBSCAN

Algorithm: DBSCAN: a density-based clustering algorithm.

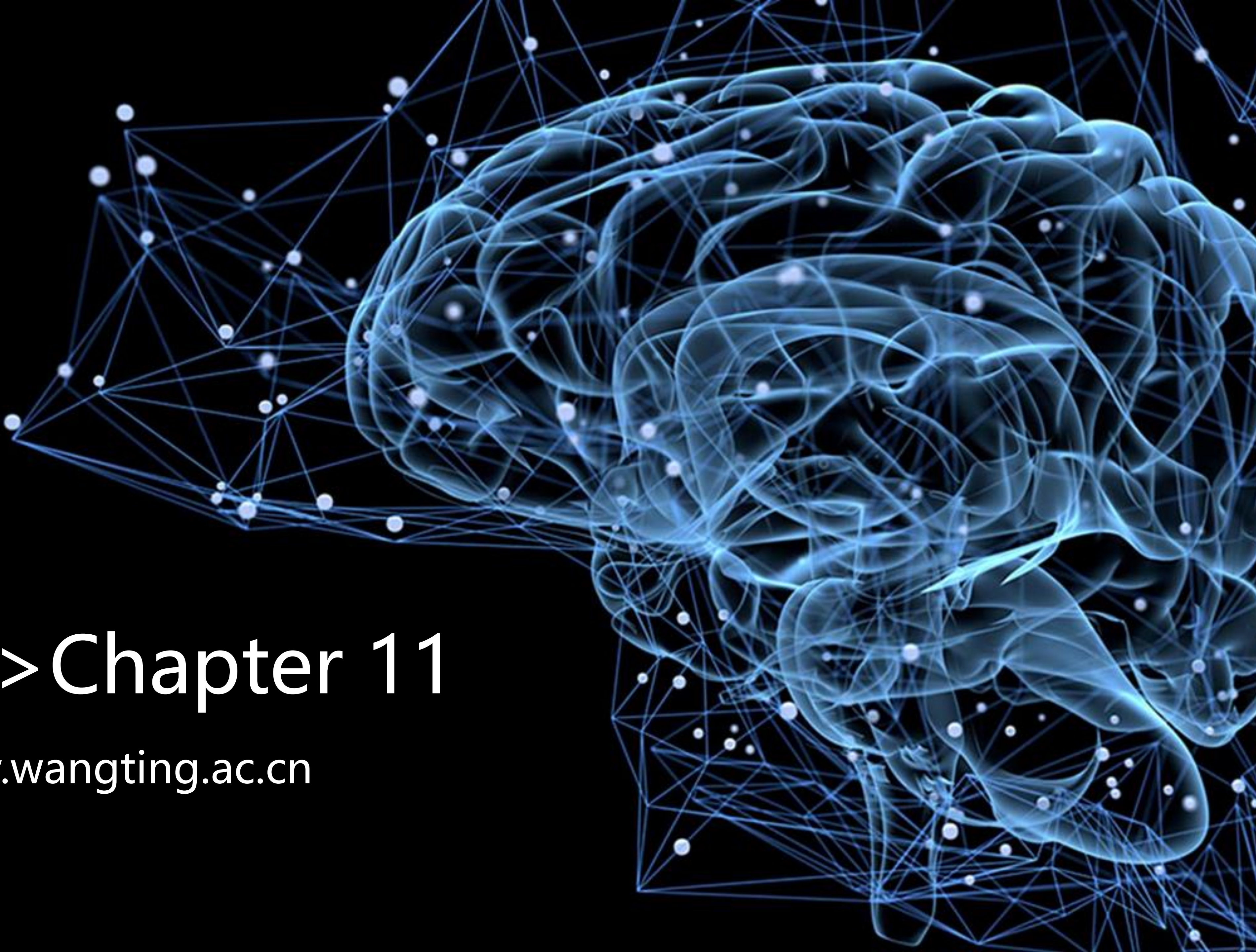
Input:

- D: a data set containing n objects,
- ϵ: the radius parameter, and
- *MinPts*: the neighborhood density threshold.

Output: A set of density-based clusters.

Method:

```
mark all objects as unvisited;
do
      randomly select an unvisited object p;
      mark p as visited;
      if the ε-neighborhood of p has at least MinPts objects
            create a new cluster C, and add p to C;
            let N be the set of objects in the ε-neighborhood of p;
            for each point p′ in N
                  if p′ is unvisited
                        mark p′ as visited;
                        if the ε-neighborhood of p′ has at least MinPts points,
                        add those points to N;
                  if p′ is not yet a member of any cluster, add p′ to C;
            end for
            output C;
      else mark p as noise;
until no object is unvisited;
```



Next>>Chapter 11

www.wangting.ac.cn