Clustering: Concepts, Methods and Evaluation

Richard Liu

School of Mathematics, XMU

August 1, 2020

Content

Source

- Pang Ning Tan, *Introduction to Data Mining*

Section 1

Introduction

Introduction

- CLustering: An unsupervised classification method.
- In this chapter, we will introduce only exclusive, not overlapping or fuzzy methods. Only complete not partial clustering algorithms.
- K-means and DBSCAN are two classical clustering methods. Although many other methods such as BIRCH and CURE are proposed, we will not introduce them due to time limits.

Section 2

K-means

K-means: General Procedure

Algorithm 8.1 Basic K-means algorithm.

Select K points as initial centroids.
repeat
    Form K clusters by assigning each point to its closest centroid.
    Recompute the centroid of each cluster.
until Centroids do not change.

K-means: Graph Illustration

(a) Iteration 1. (b) Iteration 2. (c) Iteration 3. (d) Iteration 4.

- How to select points?
- The distance metric could be changed, it depends on the objective function.

Why Centroid?

- In fact, if the objective function is

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2$$

where c_i is the centroid and x means the data points. By F.O.C we have

$$\frac{\partial}{\partial c_k} SSE = \frac{\partial}{\partial c_k} \sum_{i=1}^{K} \sum_{x \in C_i} (c_i - x)^2 = \sum_{x \in C_k} 2(c_k - x) = 0$$

This means $x = \frac{1}{m_k} \sum_{x \in C_k} x_k$, where $|C_k| = m_k$.

Why Centroid?

- However, if the objective function is

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} |c_i - x|$$

Then we have

$$\sum_{x \in C_k} 2|c_k - x| = 0$$

This means $x = median\{x \in C_k\}$, where $|C_k| = m_k$.

Bisecting K-means

Algorithm 8.2 Bisecting K-means algorithm.

Initialize the list of clusters to contain the cluster consisting of all points.
repeat
    Remove a cluster from the list of clusters.
    {Perform several "trial" bisections of the chosen cluster.}
    for $i = 1$ to *number of trials* **do**
        Bisect the selected cluster using basic K-means.
    end for
    Select the two clusters from the bisection with the lowest total SSE.
    Add these two clusters to the list of clusters.
until Until the list of clusters contains K clusters.

Pros and Cons

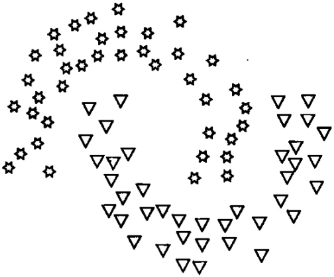
Pros:

- Efficient.
- Could be applied into different cases such as document data.

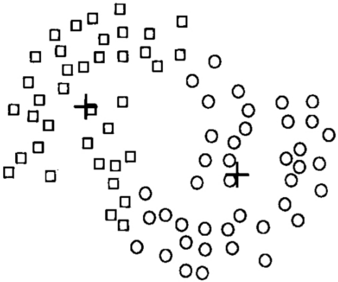
Cons:

- Not suitable for non-globular data, unequal size data and unequal density data.
- Hard to select points by random initialization and "farthest strategy".
- Hard to select K.

Non-globular Data

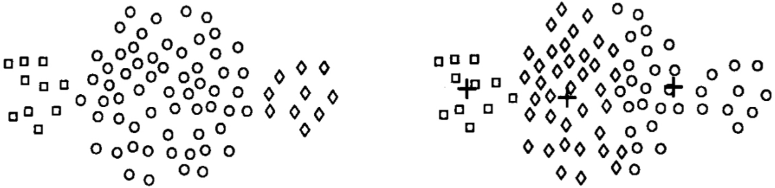


(a) Original points.



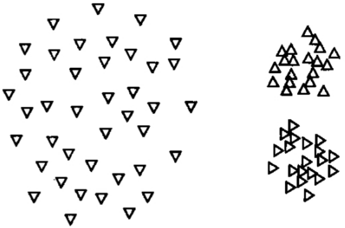
(b) Two K-means clusters.

Unequal Size Data

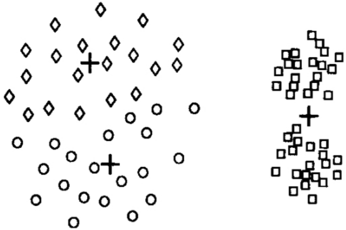


(a) Original points. (b) Three K-means clusters.

Unequal Density Data



(a) Original points.



(b) Three K-means clusters.

Section 3

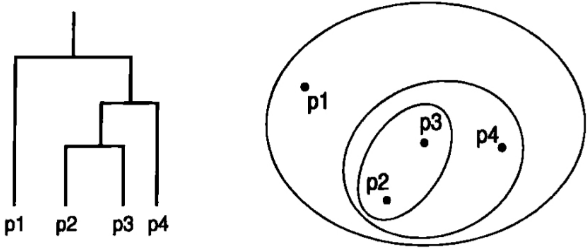
Agglomerative Hierarchical Clustering

Introduction

- Hierarchical Clustering: Another kind of clustering techniques.
  - Agglomerative: Merge, merge and merge.
  - Divisive: Split, split and split.
- We will draw a dendrogram to describe the result.

Dendrogram and Nested Cluster Diagram

516 Chapter 8 Cluster Analysis: Basic Concepts and Algorithms



(a) Dendrogram. (b) Nested cluster diagram.

Agglomerative Hierarchical Clustering: General Procedure

Algorithm 8.3 Basic agglomerative hierarchical clustering algorithm.

```
Compute the proximity matrix, if necessary.
repeat
    Merge the closest two clusters.
    Update the proximity matrix to reflect the proximity between the new
    cluster and the original clusters.
until Only one cluster remains.
```

- How to compute distance?
- What is proximity?

Example Data: *xy* Coordinates of points

Point	x Coordinate	y Coordinate
p_1	0.40	0.53
p_2	0.22	0.38
p_3	0.35	0.32
p_4	0.26	0.19
p_5	0.08	0.41
p_6	0.45	0.30

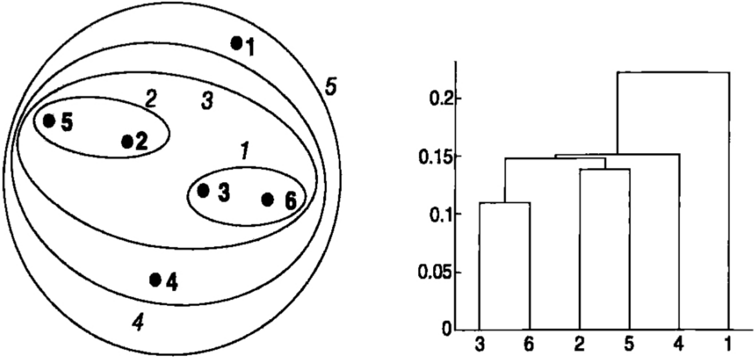
Example Data: Distance Matrix

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0.00	0.24	0.22	0.37	0.34	0.23
p_2	0.24	0.00	0.15	0.20	0.14	0.25
p_3	0.22	0.15	0.00	0.15	0.28	0.11
p_4	0.37	0.20	0.15	0.00	0.29	0.22
p_5	0.34	0.14	0.28	0.29	0.00	0.39
p_6	0.23	0.25	0.11	0.22	0.39	0.00

Proximity Measurements

- MIN (Single Link), MAX (Complete Link), Group Average and Wald's Method (not explain here).
- Differences?

Single Link

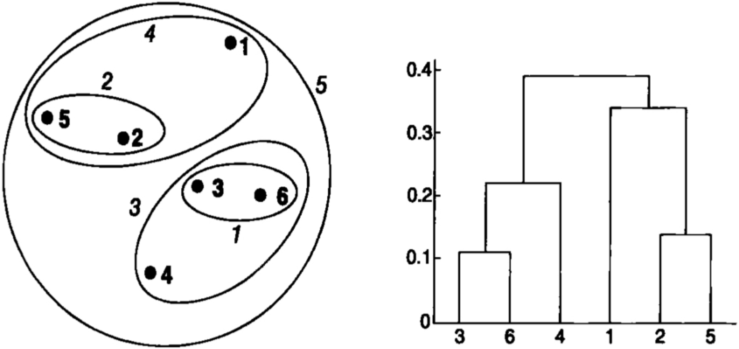


(a) Single link clustering.

(b) Single link dendrogram.

Figure 8.16. Single link clustering of the six points shown in Figure 8.15.

Complete Link

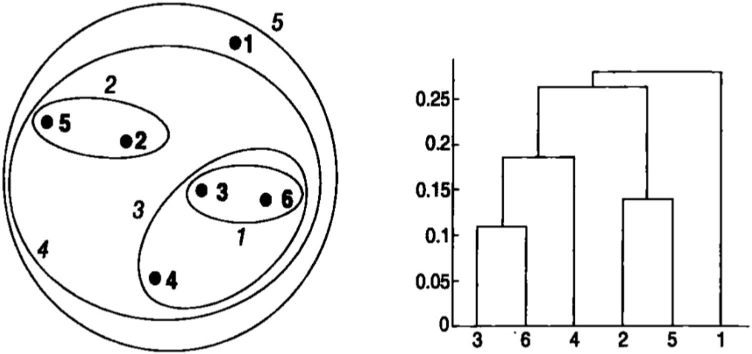


(a) Complete link clustering.

(b) Complete link dendrogram.

Figure 8.17. Complete link clustering of the six points shown in Figure 8.15.

Group Average



(a) Group average clustering.

(b) Group average dendrogram.

Figure 8.18. Group average clustering of the six points shown in Figure 8.15.

Pros and Cons

Pros:

- Easy to understand the hierarchy.

Cons:

- Not so efficient.
- No global objective function.
- The merge is final.

By the way, we usually initialize with K-means to prevent some noises, outliers and so on.

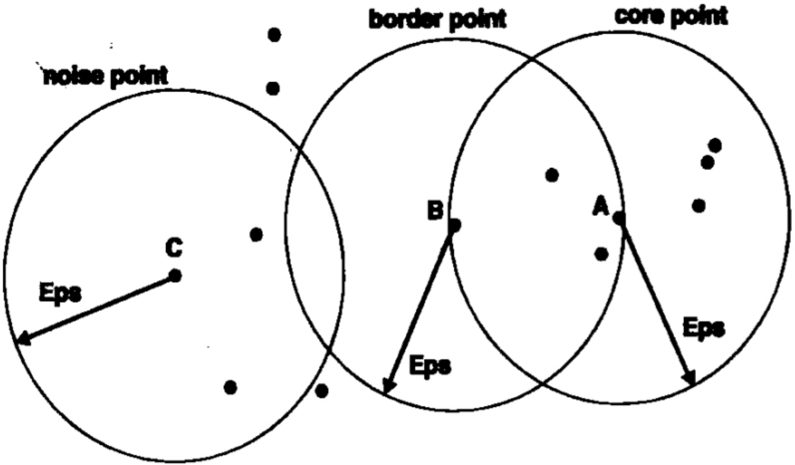
Section 4

Density-Based Clustering: DBSCAN

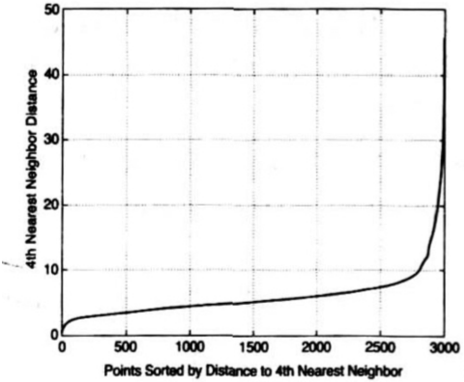
Introduction

- Density-Based Clustering Method
- Three concepts: core, border and noise points.
- Two parameters: min distance and number of points within a range.
- Parameters Selection: K-dist method.
- Problem: Varying Densities, High-Dimensional Data

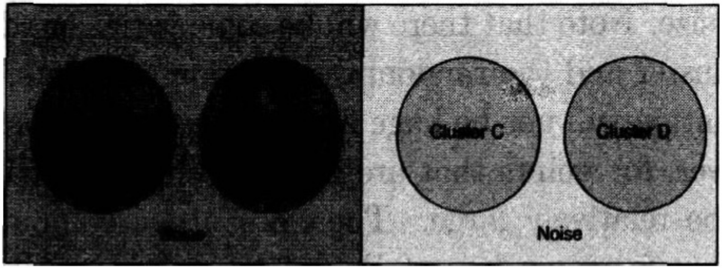
Core, Border and Noise Point



K-dist


4th Nearest Neighbor Distance
Points Sorted by Distance to 4th Nearest Neighbor
0
10
20
30
40
50
500
1000
1500
2000
2500
3000

Varying Distance



DBSCAN Result

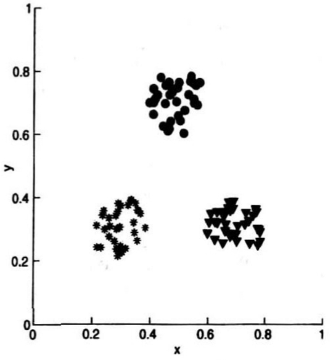
Section 5

Clustering Evaluation

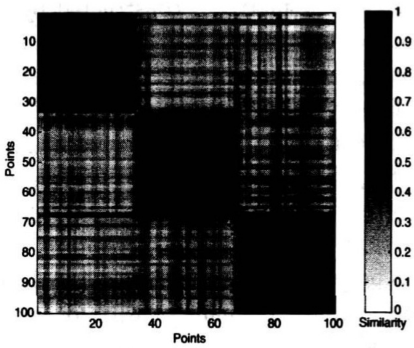
Evaluation Techniques

- Partitional Clustering: Similarity Matrix
- Hierarchical Clustering: Cophenetic Distance Matrix
- Select K in K-means: Elbow Method
- Clustering Tendency: Hopkins Statistic:
  $H = \frac{\sum_{i=1}^{p} w_i}{\sum_{i=1}^{p} w_i + \sum_{i=1}^{p} u_i}$
- CLassification-Oriented Measures: Confusion Matrix

Similarity Matrix



(a) Well-separated clusters.

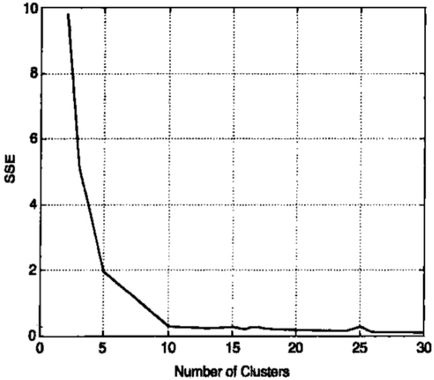


(b) Similarity matrix sorted by K-means cluster labels.

Cophenetic Distance Matrix (Single Link)

	p_1	p_2	p_3	p_4	p_5	p_6
p_1	0	0.222	0.222	0.222	0.222	0.222
p_2	0.222	0	0.148	0.151	0.139	0.148
p_3	0.222	0.148	0	0.151	0.148	0.110
p_4	0.222	0.151	0.151	0	0.151	0.151
p_5	0.222	0.139	0.148	0.151	0	0.148
p_6	0.222	0.148	0.110	0.151	0.148	0

Elbow Method



Thank you!