Clustering by fast search and find of density peaks

Alex Rodriguez, Alessandro Laio
Presenter: Honglei Zhuang

Existing Clustering Algorithms

• K-means, K-medoids
  – Data points are assigned to nearest cluster centers
  – Not applicable for nonspherical clusters
• Distribution-based
  – Assuming a generative (mixture) distribution for data
  – Requiring pre-defined distribution
Existing Clustering Algorithms

• Density-based
  – DBSCAN
    • Given a density threshold, assigns to different clusters disconnected regions of high density
    • Sensitive to the density threshold
  – Mean-Shift
    • Define a density field
    • Points converged to the same local maximum of the density field are assigned to the same clusters
    • Works only for data defined by a set of coordinates
Proposed Algorithm

• Basic Idea
  – Cluster centers are surrounded by neighbors with lower local density
  – Cluster centers are far away from other points with a higher local density

• Advantages
  – Based only on distance between data points
  – Can produce nonspherical clusters
Basic Definitions

• Local Density
  \[ \rho_i = \sum_j \chi(d_{ij} - d_c) \]
  where \( \chi(x) = 1_{\{x<0\}} \) and \( d_c \) is a given cutoff
  – Basically is the number of points closer than the cutoff to the point.

• Define
  \[ \delta_i = \min_{j: \rho_j > \rho_i} (d_{ij}) \]
  – The minimum distance to other points with a higher local density
  – Defined as \( \delta_i = \max_j (d_{ij}) \) is the density is largest
Fig. 1. The algorithm in two dimensions. (A) Point distribution. Data points are ranked in order of decreasing density. (B) Decision graph for the data in (A). Different colors correspond to different clusters.
Example

Fig. 1. The algorithm in two dimensions. (A) Point distribution. Data points are ranked in order of decreasing density. (B) Decision graph for the data in (A). Different colors correspond to different clusters.

relatively high in both $\rho$ and $\delta$
Fig. 1. The algorithm in two dimensions. (A) Point distribution. Data points are ranked in order of decreasing density. (B) Decision graph for the data in (A). Different colors correspond to different clusters.
Proposed algorithm

• After cluster centers have been found,
  – Each remaining point is assigned to the cluster of nearest neighbor of higher density
  – No need to be optimized iteratively
Reliability

• When noise exists
• Define a “border region” for each cluster where points are within cutoff distance $d_c$ from points of other clusters
• Define a highest density in the border region as the threshold density $\rho_b$
• Any points with a local density lower the threshold density is regarded as noise
Experiments

- 4,000 points drawn
Parameter Sensitivity
Experiments on Face Database
Experiments on Face Database

• Cannot clearly determine #clusters

\[ \gamma_i = \rho_i \delta_i \]
On Random Data Set

• A hint to determine the cluster centers is to calculate

\[ \gamma_i = \rho_i \delta_i \]

where usually a gap exists

• In a (uniformly) randomly distributed data set, following a power law
Summary

• Algorithm Sketch
  – Calculate local density and minimum distance to data point with higher density
  – Determine cluster centers
  – Assign data points to cluster of the closest data point with higher density

• Advantages
  – Works for nonspherical clusters
  – Only requires distance

• Drawbacks
  – Sometimes hard to determine number of clusters
Thanks

Honglei Zhuang