Paper Summary: "Clustering by fast search and find of density peaks (2014)"

Problems (WHAT & WHY)

Cluster analysis is aimed at classifying elements into categories on the basis of their similarity. Traditional methods include K-means (assigns points to its nearest center), DBSCAN. However, these methods have various limitations, either that they cannot be applied in non-spherical clusters (ex.: bivariate moon shape data), or that they are computationally costly.

Idea (HOW)

Authors of this paper think that recognizing cluster centers is the key in clustering problems, and they make a natural assumption that any cluster center has two properties:

- high local density (means there are relatively more points in the center's neighborhood)
- far from other high local density centers

Details (HOW)

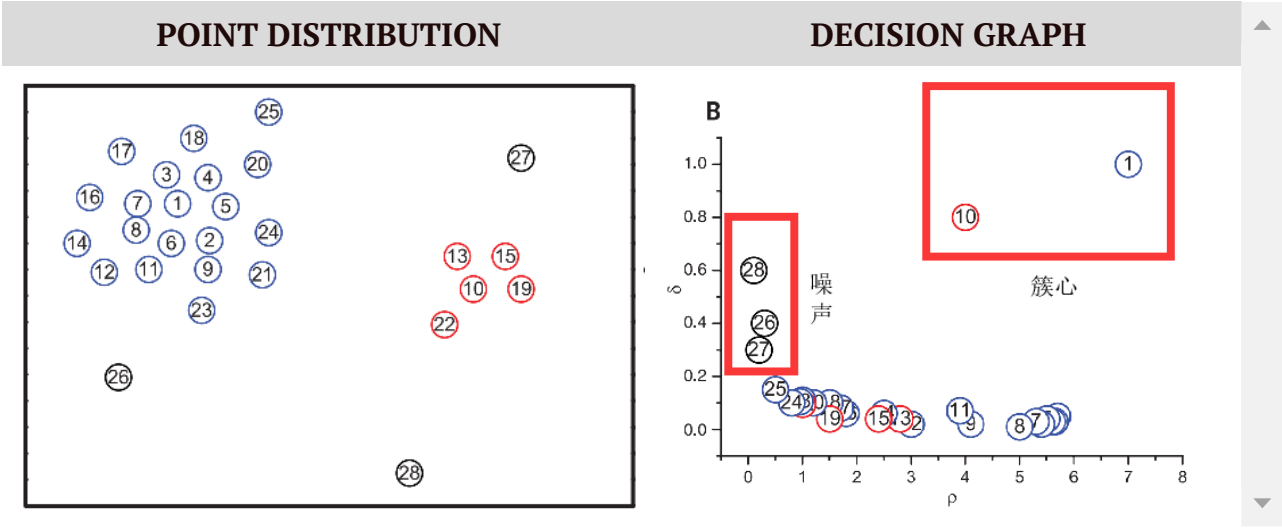
First, they define two values ρ, δ to formalize the concepts of "between-center distance" and "local density":

$$\delta_i = \min_{j:\rho_j > p_i} (d_{ij})$$
$$\rho_i = \sum_j \chi (d_{ij} - d_c)$$

here d_c is a hyperparameter, it is given by experience.

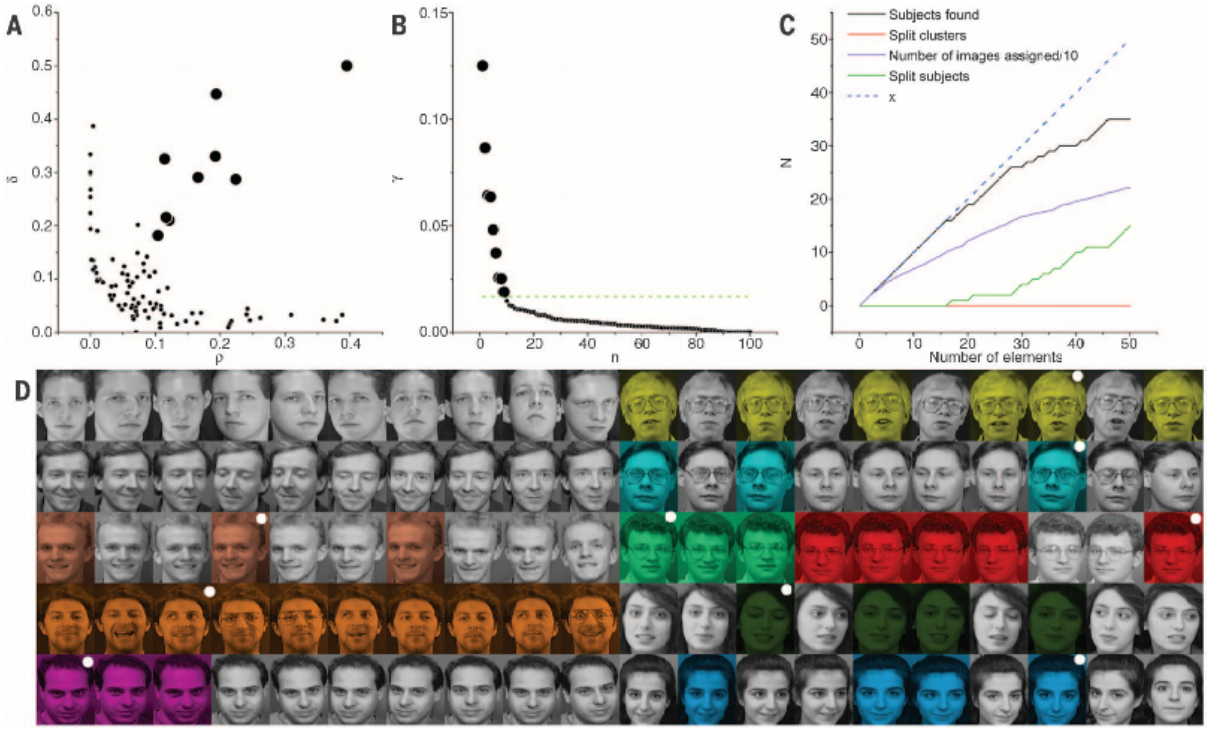
Now given a dataset, we need to do:

- step 1: calculate all ρ, δ values
- **step 2: detect cluster centers (with higher values of ρ and δ)**
- step 3: for non-center points, link them to their nearest center
- step 4: remove noise points (with small values of ρ and δ)



Example (WHERE)

Authors tested various non-spherical data and got nice performance. Specifically, in the human face clustering work, the algorithm showed high accuracy but ignored many images (took them as noise).



Summary

Pros and Cons

- The new proposed algorithm only needs distance information d_{ij} and it is fast because it does not require recursive computation like KNN, besides, it can handle non-spherical data.
- The fast computation ability of the algorithm is based on larger storage, we need all distance information, the storage cost is $O(n^2)$.

Future Work

More work can be done in how to determine a proper value for d_c, it is actually the problem of how should we define similarity, the problem may be solved by the neural network.

Reference

论文精读报告-Clustering by fast search and find of density peaks