Paper One

Clustering by fast search and find of density peaks

Motivation

Traditional clustering method can be divided into two categories: distance-based method and density-based method. The former one is not able to detect arbitrary-shaped clusters and the latter one is usually computationally expensive. Therefore, the paper points out a new way to do the clustering work combining thoughts of distance-based method and thoughts of density-based method.

Related Definition

- Density: $\rho_i = \sum_j \chi(d_{ij} - d_c)$, where $\chi(x) = \begin{cases} 1, & x < 0; \\ 0, & x \geq 0, \end{cases}$.

That is to say, “density” of a point_i in this paper means the number of other points whose distance from the point_i is less than dc.

- Distance: $\delta_i = \min\limits_{j:\rho_j > \rho_i} (d_{ij})$. In particular, the distance of the point with highest density: $\delta_i = \max_j(d_{ij})$

That is to say, “distance” of a point_i in this paper means that the minimum distance between point_i and someone with higher density.

Main Algorithm

1. Firstly compute all the distance d_{ji}

   (actually, computing the half is enough since the distance $d_{ji} = d_{ij}$)

2. Choose the parameter d_c

   (This paper introduces a rule of thumb to determine d_c that the number of each point's neighbors is about 1%~2% of the total is proper. And the neighbors of a point_i means those points whose distance is less than d_c)

3. Compute the density ρ and the distance δ for each point

   (The computational formula of ρ、δ are listed in the above definition)

4. Then normalize ρ and δ to compute $\gamma = \rho \cdot \delta$, from which we can select the cluster centers.

   (According to this paper, we can choose the cluster centers by plotting the γ-value in a descending order. In this step, we can not only determine the number of clusters but the cluster centers as well.)

5. Next we classify the other points into several clusters we choose in the above step.

   (The cluster of a point is determined by its δ. It should be the same cluster as $\arg\delta$)

6. In the last step, we should distinguish the cluster halo since we have categorized all the points.

   (Cluster halo is defined as: In its range of d_c, there exists some points which belong to other cluster and the average density $\overline{\rho} = \frac{1}{2}(\rho_i + \rho_j)$ is larger than its own density ρ_i)

Contribution

The authors give us a new idea to do the clustering work based on both density and distance. The algorithm is capable to recognize the clusters with arbitrary shape while it is not computationally costly. What's more, the number of clusters can be determined automatically and outliers can be found as well.

A little Improvement

After reading this paper, I also read some reports and blogs about the paper. Though there is some criticism on this article, I think the paper gives us an entirely new idea on data clustering. But some details on this algorithm can be improved. Here are two little improvement on its definition about density and distance:

- Density: $\rho_i = \sum_j e^{-(\frac{d_{ij}}{d_c})^2}$ -- Gaussian kernel

Still, the smaller distance is, the bigger density is. But it will have fewer collisions since it becomes continuous value instead of discrete value.

- Distance: According to old definition, when two points both have the same highest density, their distance calculation formula is $\delta_i = \max_j (d_{ji})$.

However, if the two points are actually close, they will be divied into two clusters as cluster centers since they all have relatively high density and high distance. That's the problem. We can improve it by just ignore the second point with highest density, which means only one point with highest density can be calculated in this way $\delta_i = \max_j (d_{ji})$.

Paper two

Robust continuous clustering

Motivation

Existing clustering methods perform badly in massive datasets with high dimensions and usually need to turn parameters to adapt different domains. Therefore, the paper tries to give us a new algorithm that can scale to high dimensions and massive datasets, and achieve high accuracy across many different domains as well.

Related definition and theory

Objective function:

$$\mathbf{C}(\mathbf{U},\mathbb{L}) = \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \frac{\lambda}{2}\sum_{(p,q)\in\mathcal{E}} w_{p,q}\left(l_{p,q}\|\mathbf{u}_p - \mathbf{u}_q\|_2^2 + \Psi(l_{p,q})\right).$$

Where u is the representatives of data point, l_{pq} is the line process or connection activity and they are both updated iteratively by linear least-squares solver. Also, $\Psi(l_{pq})$ satisfy the relations as the following:

$$\begin{cases} \Psi(l_{pq}) \xrightarrow{l_{pq}\to 1} 0 \\ \Psi(l_{pq}) \xrightarrow{l_{pq}\to 0} 1 \end{cases}$$

When U are fixed, the optimal value of l_{pq} is as following according to the paper: $l_{p,q} = \left(\frac{\mu}{\mu + \|\mathbf{u}_p - \mathbf{u}_q\|_2^2}\right)^2.$, where μ is a parameter and must satisfy $\mu \gg \max\|x_p - x_q\|_2^2$.

When L are fixed, the optimal formula is as following according to the paper:

$$\arg\min \frac{1}{2}\|\mathbf{X}-\mathbf{U}\|_F^2 + \frac{\lambda}{2}\sum_{(p,q)\in\mathcal{E}} w_{p,q} l_{p,q} \|\mathbf{U}(\mathbf{e}_p-\mathbf{e}_q)\|_2^2,$$

, where λ is a parameter as well, and usually initialize it equal to $\frac{\chi}{\|A\|_2}$.

Convergence conditions : The difference of objective function values between this iteration and the last one iteration is smaller than a small parameter which means the iteration process has converged, or the number of iterations has reach the maximum which means we have to stop.

Rough Process of the Algorithm

```
1. Build connectivity structure $\varepsilon$
2. Compute $\chi, w_{pq}, \delta$
3. Initialize $u_i = x_i, l_{pq} = 1, \mu, \lambda$
4. While $|C^t - C^{t-1}| < bound$ or $t < \max iterations$
        Update $l_{pq}$ and A
        Update U
        Update $\lambda = \frac{\chi}{\|A\|_2}$, $\mu = \max(\frac{\mu}{2}, \frac{\delta}{2})$ after each four iterations
5. Construct new graph G with $f_{pq} = 1$ if $\|u_p - u_q\|_2 < \delta$
6. Output clusters on the basis of the connected parts of G
```

Contribution

Honestly speaking, I did not clearly figure out the details of algorithm and its mathematical theories. But experimental results show that RCC algorithm can deal with massive dataset with high dimensions, adapting to different domains. What’s more, we have no need to choose the number of clusters in advance when using RCC algorithm. In the paper, authors also update the algorithm to do the dimensionality reduction work.