|  |  |  |
| --- | --- | --- |
|  | **hierarchical single link** | **k-means** |
| a |  |  |
| b |  |  |
| c |  |  |

**Spectral Clustering**

1. Data\_a: #5=#8=#2>#1









1. Data\_b: #5>#2=#1>#8









1. Data\_c: #5>#8>#2>#1









**Discussion:**

1. The performance of the three clustering method:

For data base a, k-means and spectral clustering perform better than single linkage.

For data base b, single linage is better than spectral clustering, and spectral clustering better than k-means.

For data base c, spectral clustering and single linkage is better than k-means.

* There’s not a method that is consistently better than others.
* Single linkage has better performance when the distance between clusters are similar, since it merges in each step the two clusters whose two closest members have the smallest distance. So, for database like (a), it merges the two cluster together which should be separate.
* K-means performs better when the pattern is convex, has similar density and size. Since it is based on the calculation of center, and concave patterns’ center is outside the pattern itself, so it may not be accurate in that way. For database like (c), k-means may think all the points around center are from a same cluster but they’re actually not.
* Spectral clustering performs better when the data has good connectivity. Since it is based on the distance matrix and its Laplacian matrix. So when the data are not closely connected or has several noise data points (such as that in b), it may give out a less satisfied clustering.

1. For database (a) and (b), it is possible to use linear classification method to divide the data points using support vector machine. However, for database (c), it seems impossible to do so.

* For (a), we should split each cluster on the outer side one by one, and find its support vectors to other four clusters. Then we end up with four lines which separate the whole domain as shown below. However, it is much more complicated and time consuming than the three methods mentioned in this question.



* For (b), it is possible to split the data using support vector machines and it will end up like this.



For (c), since the pattern is not linear separable, support vector machine cannot accurately separate these data points. We cannot find an appropriate margin value on convex patterns.