CSCI3220 2018-19 First Term Assignment 5

I declare that the assignment here submitted is original except for source material explicitly acknowledged, and that the same or closely related material has not been previously submitted for another course. I also acknowledge that I am aware of University policy and regulations on honesty in academic work, and of the disciplinary guidelines and procedures applicable to breaches of such policy and regulations, as contained in the following websites.

University Guideline on Academic Honesty: http://www.cuhk.edu.hk/policy/academichonesty/

Student Name: ZHANG Chongzhi
Student ID : 1155077072

1.

(a).

X	s_1	s_2	s_3	s_4	s_5	s_6
g_1	4	6	5	1	2	8
g_2	1	4	6	8	1	3
g_3	2	9	3	5	1	6
g_4	8	5	2	1	3	4
g_5	7	1	1	3	2	9

$d(g_1, g_2) = \frac{\sum_{k=1}^{l} (x_{ik} - x_{jk})^2}{6} = \frac{9+4+1+49+1+25}{6} = 14.8$

As the same method:

$d(g_1, g_3) = 6.3$ $d(g_1, g_4) = 7.2$ $d(g_1, g_5) = 9.2$

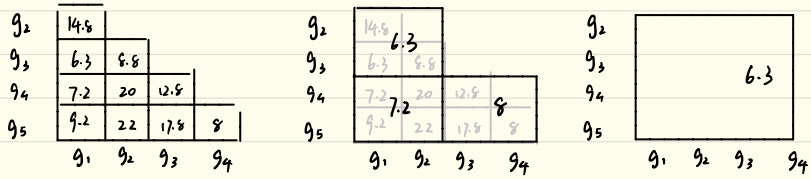
$d(g_2, g_3) = 8.8$ $d(g_2, g_4) = 20$ $d(g_2, g_5) = 22$

$d(g_3, g_4) = 12.8$ $d(g_3, g_5) = 17.8$ $d(g_4, g_5) = 8$

Produce the distance matrix.

d	g_1	g_2	g_3	g_4	g_5
g_1	0	14.8	6.3	7.2	9.2
g_2	14.8	0	8.8	20	22
g_3	6.3	8.8	0	12.8	17.8
g_4	7.2	20	12.8	0	8
g_5	9.2	22	17.8	8	0

(b). Quad tree :



(C) From the initial quad tree, we knows that g_1 and g_3 are merged first.

1). single-link : $d(\{g_1, g_3\}, \{g_k\}) = \min(d(\{g_1\}, \{g_k\}), d(\{g_3\}, \{g_k\}))$

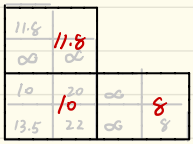
g_2	g_4	g_5
8.8	7.2	9.2

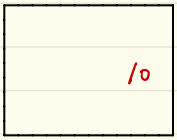
	g_2	g_4	g_5
g_2	8.8		
g_4	7.2	20	∞
g_5	9.2	22	∞	8
	$g_1 g_3$	g_2		g_4

2). Average-link : $d(\{g_1, g_3\}, \{g_k\}) = \dfrac{d(\{g_1\}, \{g_k\}) \cdot 1 \cdot 1 + d(\{g_3\}, \{g_k\} \cdot 1 \cdot 1)}{2 \times 1}$

g_2	g_4	g_5
11.8	10.0	13.5

	g_2	g_4	g_5
g_2	11.8		
g_4	10	20	∞
g_5	13.5	22	∞	8
	$g_1 g_3$	g_2		g_4

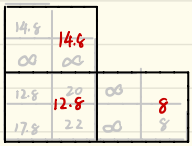


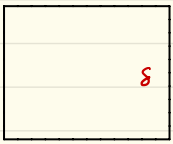


3) Complete-link : $d(\{g_1, g_3\}, \{g_k\}) = \max\{d(\{g_1\}, \{g_k\}), d(\{g_3\}, \{g_k\})\}$

g_2	g_4	g_5
14.8	12.8	17.8

	g_2	g_4	g_5
g_2	14.8		
g_4	12.8	20	∞
g_5	17.8	22	∞	8
	$g_1 g_3$	g_2		g_4





3. (a). Every time when re-determining representative, the cluster that contain the outlier will generate a representative dragged to the outlier point. Thus next time, the cluster that contain that point may have less points, which means the representative will more closer to the outlier point.

Therefore, if the outlier point is far enough, it will become a cluster itself and all others points merged to the other cluster.

(b). The column has a much larger magnitude, which means that the representative will be very close to the center of that column. Each times do the iteration, the cluster has a trend to have less point outside that column.

Therefore, the column and the points that close to its center will be merged into one cluster.

(c). Advantage: It uses a ratio rather than a real distance, the difference will be increase for close point and decrease from far-away points. It will generate clusters with more "relatively-closed" points.

Disadvantage: More calculation required.

(d). First let's just make the $k \leftarrow 2$, and make the k increase by the variance.

Each time when we get k clusters,

for each cluster, do

check the variance of all the points in that cluster, if the variance is greater than some given value threshold, let the cluster split: $k \leftarrow k+1$.

if all the cluster has a variance less than threshold,

return k.