

# Evaluation of the Optimal Clustering Algorithm and the Linear Assignment Clustering Algorithm

**Cheng-Feng Sze**

Dept. of Electrical Engineering  
University of Maryland  
College Park, MD 20742

January, 2000

We evaluate the two clustering algorithms based on the correctness and speed of the clustering algorithms.

## I. Correctness of the clustering result

A clustering algorithm is to cluster similar objects together. There are some reasonable cost functions for comparing the correctness of the clustering algorithms.

### 1. Total average distance of points in clusters

For a cluster  $I = \{i_1, \dots, i_n\}$ , the average distance of points in the cluster  $I$  can be defined as

$$S(I) = \frac{2}{n(n-1)} \sum_{k=1}^{n-1} \sum_{h=k+1}^n d(i_k, i_h).$$

Note that  $\frac{n(n-1)}{2}$  is the total number of the summation.

For clusters  $\{I_1, \dots, I_N\}$ , the total average distance of points can be defined as

$$T(I_1, \dots, I_N) = \frac{1}{\sum_{k=1}^N \rho(I_k)} \sum_{k=1}^N S(I_k) \rho(I_k),$$

where  $\rho(I)$  is the number of points in the cluster  $I$ .

Remark:  $T(I_1, \dots, I_N)$  measures the average distance of points in all the clusters. A smaller value of  $T(I_1, \dots, I_N)$  represents a better clustering result.

For the points in Figure 4, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
$S(I_1)$	0.5235	1.3167
$S(I_2)$	1.6410	1.5262
$T(I_1, I_2)$	0.9593	1.3502

Therefore, the optimal clustering algorithm gave a better result.

For the points in Figure 5, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
$S(I_1)$	0.4432	1.0868
$S(I_2)$	1.6253	1.2215
$T(I_1, I_2)$	0.9751	1.1218

Therefore, the optimal clustering algorithm gave better result.

## 2. Average of maximum distance of points in clusters

for a cluster  $I$ , let  $M(I)$  be the maximum distance of points in cluster  $I$ . Then, for clusters  $\{I_1, \dots, I_N\}$ , the average of maximum distance of points can be defined as

$$A(I_1, \dots, I_N) = \frac{1}{N} \sum_{k=1}^N M(I_k).$$

For the points in Figure 4, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
$M(I_1)$	1.9765	3.4416
$M(I_2)$	4.7409	4.2090
$A(I_1, I_2)$	3.3587	3.8253

Therefore, the optimal clustering algorithm gave better result.

For the points in Figure 5, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
$M(I_1)$	1.5445	2.6526
$M(I_2)$	4.7409	4.3233
$A(I_1, I_2)$	3.1427	3.4879

Therefore, the optimal clustering algorithm gave better result.

## 3. Average representation error

In coding, we want to use the centers of clusters to represent those clusters for data reduction. Therefore, we can calculate the average mean square error for the representations.

For the points in Figure 4, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
Center of $I_1$	[-0.4176 -0.1410]	[0.227 1.0969]
Center of $I_2$	[1.4565 0.0513]	[0.811 -0.2371]
M.S.E./points	0.6050	1.1504

Note that the actual centers of the two clusters should be [0 0] and [1.5 0].

Based on the average m.s.e. and the center estimation, the optimal clustering algorithm gave better result.

For the points in Figure 5, we have the following table.

	optimal clustering algorithm	Linear assignment clustering
Center of $I_1$	[-0.4176 -0.1410]	[-0.6954 -0.6824]
Center of $I_2$	[1.8664 0.0513]	[1.5627 0.2077]
M.S.E./points	0.6544	0.7202

Note that the actual centers of the two clusters should be [0 0] and [2 0].

Based on the average m.s.e. and the center estimation, the optimal clustering algorithm gave better result.

4. A cluster algorithm is not only to cluster similar objects together, but also to explore the structures between clusters. Note that the points in Figure 4 & 5 are generated from two sources: one has very small variance and the other one has much larger variance. Therefore, a good clustering scheme should have the ability to explore this kind of phenomenon. That is, the points that close to each other should be in different cluster with the points that scatter. Based on this point, the optimal clustering scheme did a better job. (Note that from Fig 4(b) & 5(a) the linear assignment clustering algorithm did not well separate the points of the two sources, especially for the points in Fig. 4(b).)

## II. Speed of the algorithm

Based on the speed, the linear assignment clustering algorithm is much faster than the optimal clustering algorithm. However, due to poor clustering performance, the linear assignment clustering algorithm has limited applications (since the result is unacceptable). On the other hand, the optimal clustering algorithm can be applied to those applications in which the calculation time is not crucial.