Your Name: ______Yanfang Wang______

(This is an INDIVIDUAL assignment)

CSC 7442: Data Mining and Knowledge Discovery from Databases

Professor E. Triantaphyllou
Louisiana State University
Department of Computer Science
Fall 2019

Today's date:	**Thursday, September 26, 2019**
Due date:	**Tuesday, October 8, 2019. By 10:00 pm of the day via email to trianta@csc.lsu.edu (note this is NOT my PAWS address. This is my CSC address)**

Computer Assignment #1: (Maximum = 200 points)

MAIN GOALS: **To explore some fundamental issues related to data and the concept of clustering**

Description of the computing assignment on clustering

See the attached dataset in excel format. It describes a simple dataset in two dimensions. Therefore, you can plot it easily. We want to determine different clusters from this dataset. We do not know the number of clusters or anything else. Use the simple **K-means method** and also the method based on **bisecting of clusters** and the method based on **dispersing of clusters** (for the later method start with K+3 clusters if you want to have K clusters at the end).

Present your results in the form of tables and/or graphs. Compare all your results and comment on which one you think is the best clustering solution.

Submit your computer programs too and be prepared to run them on a PC if you are asked to do so. Use any programming language you wish. Make sure you document your coding well.

Attach this form on front of your answers

1. Dataset

The original data is given below in Fig. 1, the following sections are the clustering results from different clustering methods based on K-Means, general K-Means, Bisecting clustering, and dispersing of clusters, respectively.

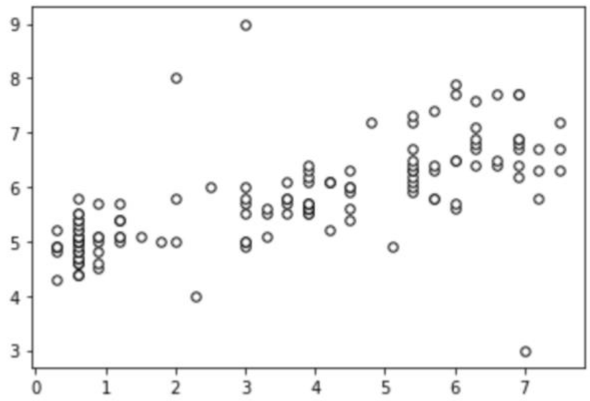

Fig. 1 Dataset

2. K-Means method

The table below Table 1 is the sum of squared errors using Euclidean distances. The SSEs are calculated based on number of clusters in the whole dataset. The way to determine the optimum number of clusters for the best solution is called Elbow method. Fig. 2 shows the SSEs as function of number of clusters. From this plot, we can see the global SSE decreases as number of clusters increases. I plotted the cases near the inflection point, K = 2, 3, 4. The final clustering solutions are shown in Fig. 3(a), 3(b), 3(c).

Table 1

K	1	2	3	4	5	6	7	8	9	10
SSE	319.918	158.28	99.049	86.63	81.178	77.223	68.942	67.556	59.249	57.986

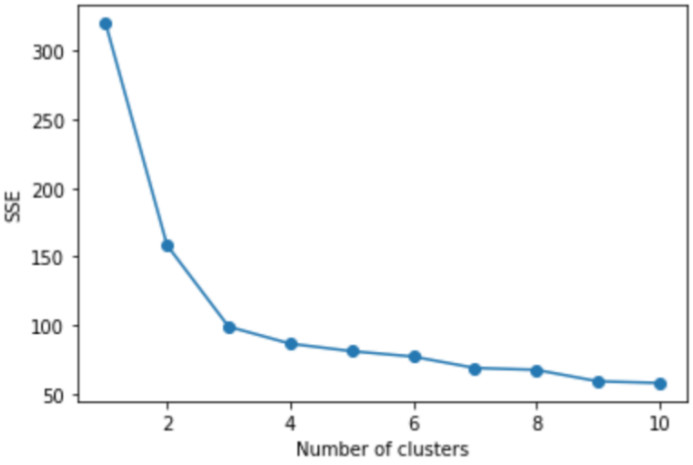

Fig. 2 Determine the best solution for clustering using K-Means

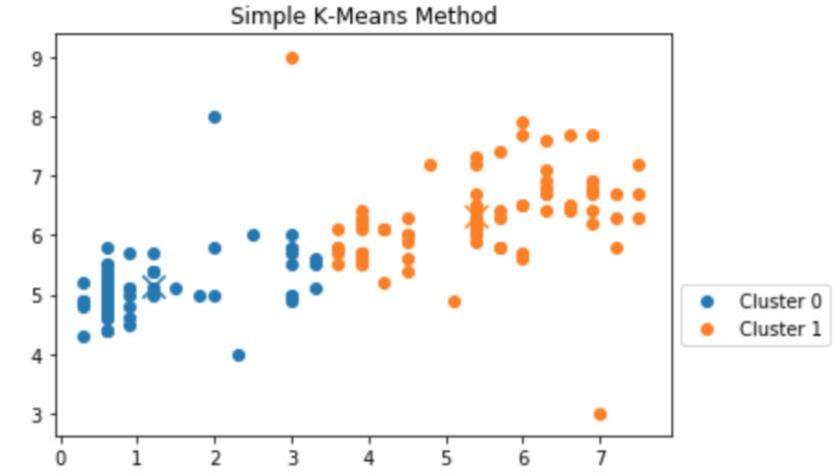

Fig. 3(a) The best solution (k = 2) using K-Means method

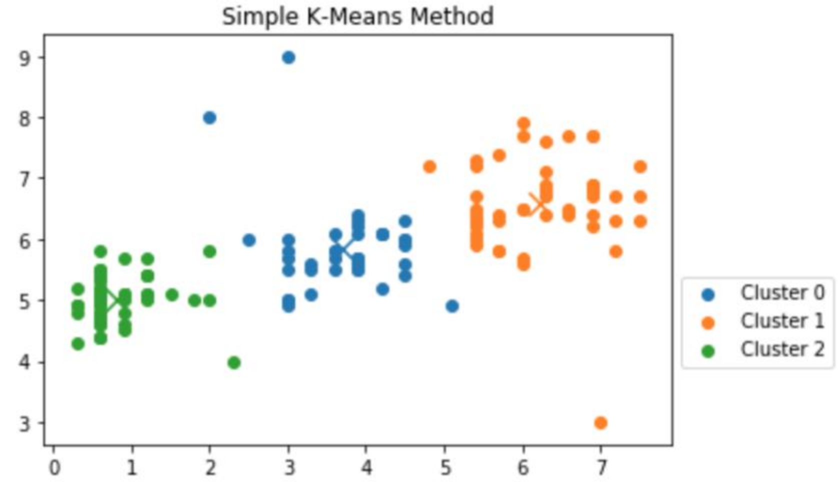

Fig. 3(b) The best solution (k = 3) using K-Means method

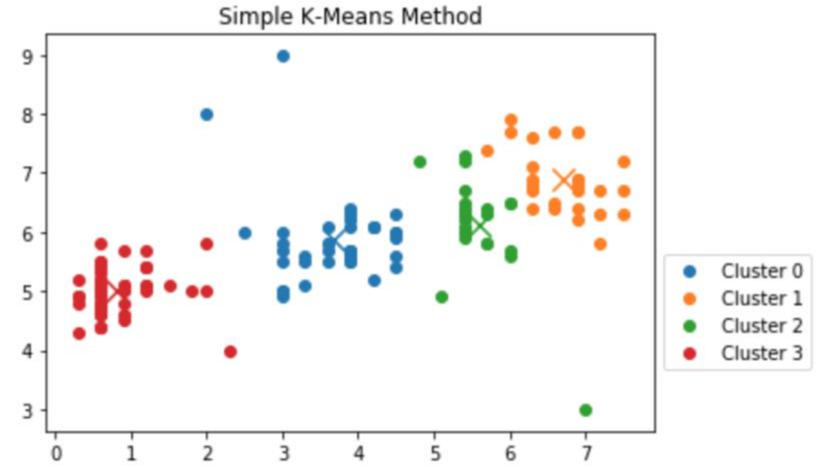

Fig. 3(c) The best solution (k = 4) using K-Means method

3. Bisecting method

Below is the SSEs change with number of clusters using Bisecting clustering, shown in Fig. 4.

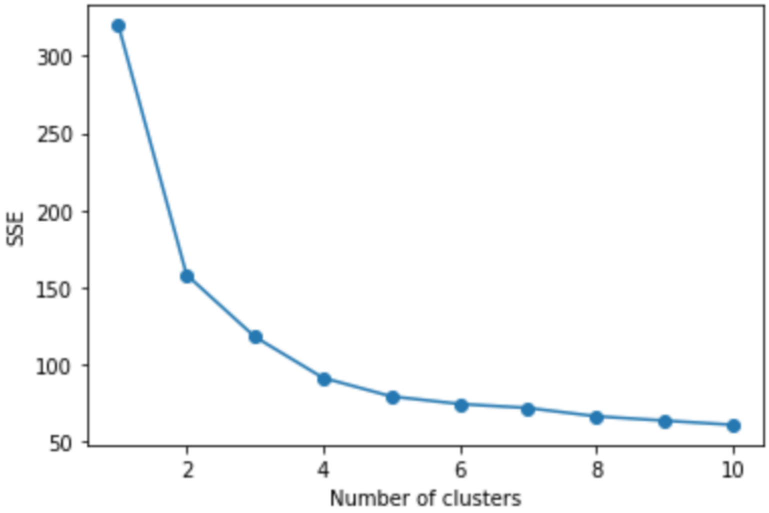

Fig. 4 Determine the best solution for clustering using Bisecting clustering

As shown in Fig. 5(a), 5(b), 5(c), I plotted the cases near the inflection point, K = 2, 3, 4.

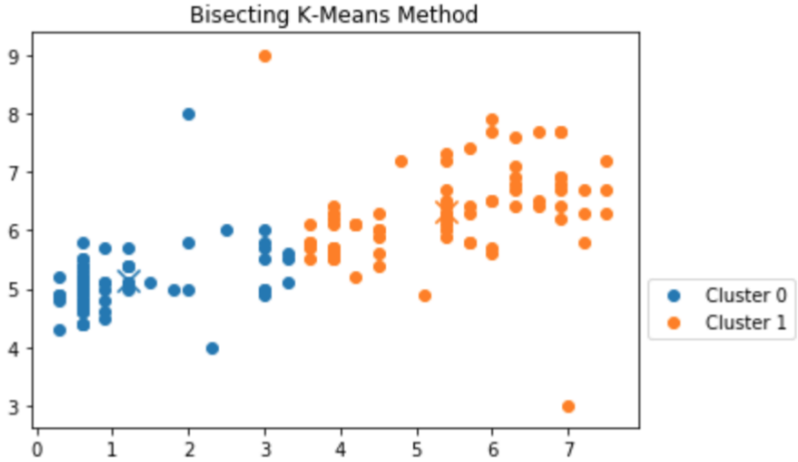

Fig. 5(a) The best solution (k = 2) using Bisecting clustering method

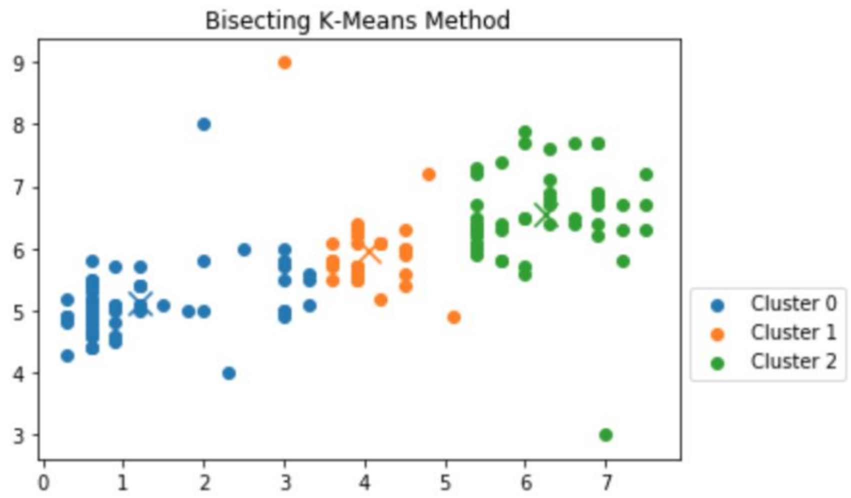

Fig. 5(b) The best solution (k = 3) using Bisecting clustering method

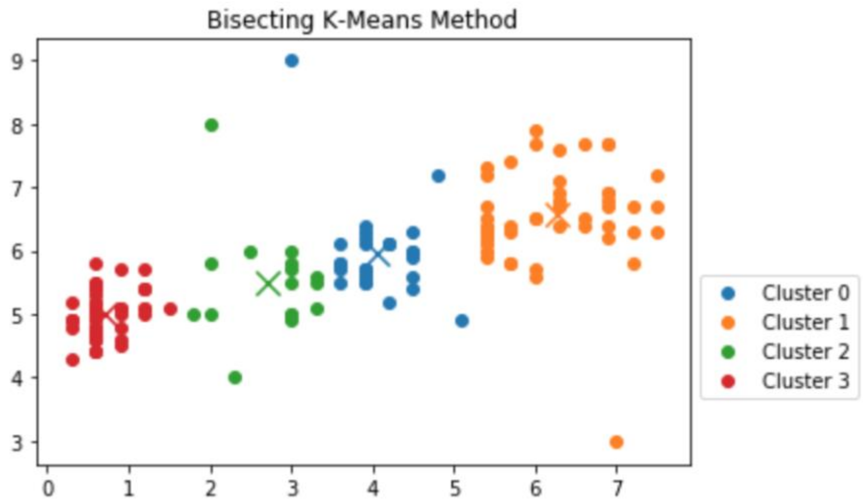

Fig. 5(c) The best solution (k = 4) using Bisecting clustering method

4. Dispersing of clustering method

Below is the SSEs change with number of clusters using Dispersing clustering, shown in Fig. 6.

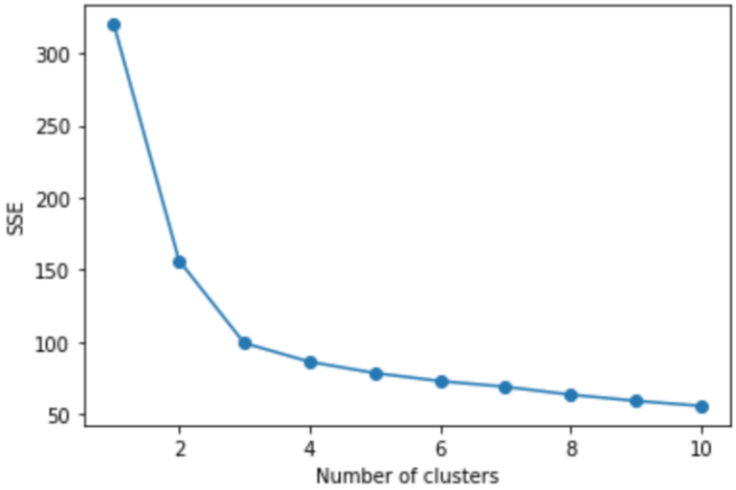

Fig. 6 Determine the best solution for clustering using Dispersing clustering

As shown in Fig. 7(a), 7(b), 7(c), I plotted the cases near the inflection point, K = 2, 3, 4.

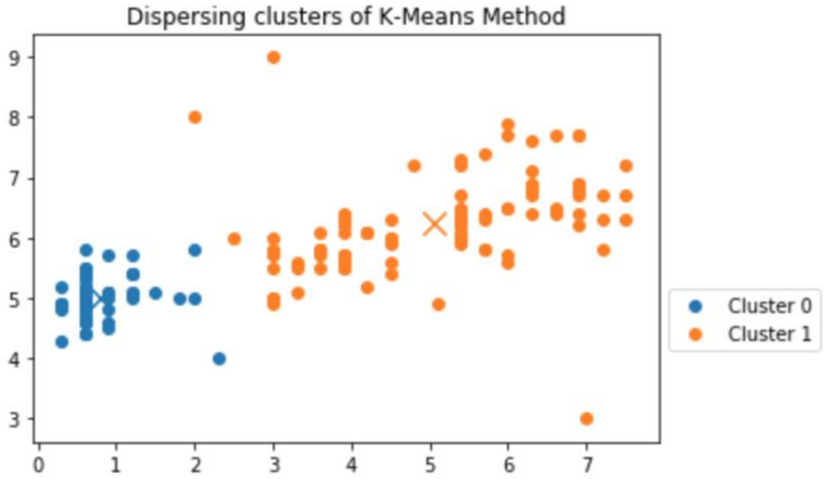

Fig. 7(a) The best solution (k = 2) using Dispersing clustering method

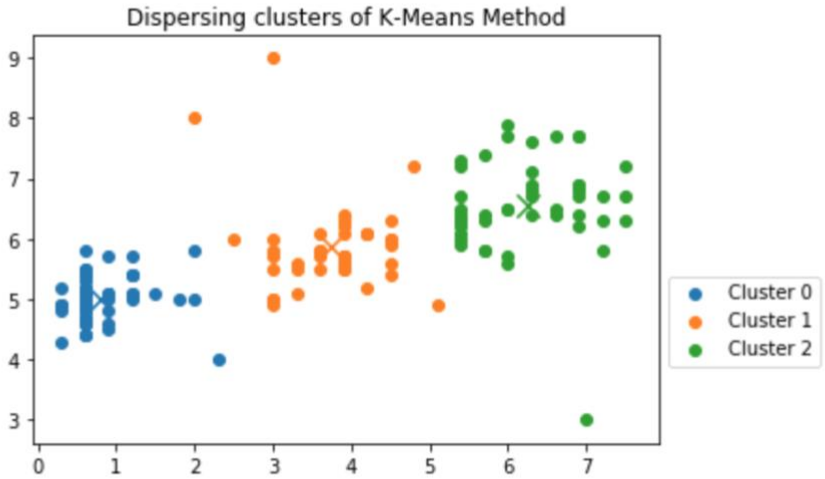

Fig. 7(b) The best solution (k = 3) using Dispersing clustering method

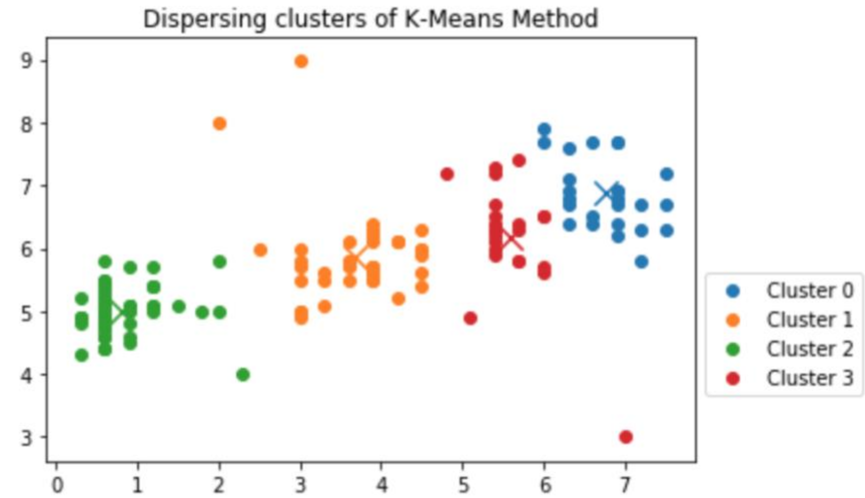

Fig. 7(c) The best solution (k = 4) using Dispersing clustering method

5. Comparisons of Different Methods

Below is the summary table of SSEs with the change of K numbers:

K	1	2	3	4	5	6	7	8	9	10
K-Means	319.918	158.28	99.049	86.63	81.178	77.223	68.942	67.556	59.249	57.986
Bisecting	319.918	158.28	118.156	91.547	79.549	74.647	72.073	66.662	63.849	61.143
Dispersing	319.918	156.5	99.374	86.432	78.636	73.074	69.130	63.6816	59.38	55.909

Based on the magnitude of SSEs values, Bisecting method is not better than the other two clustering methods. Dispersing method gives a little bit lower SSE values than K-Means methods, from K = 2 to K = 10.