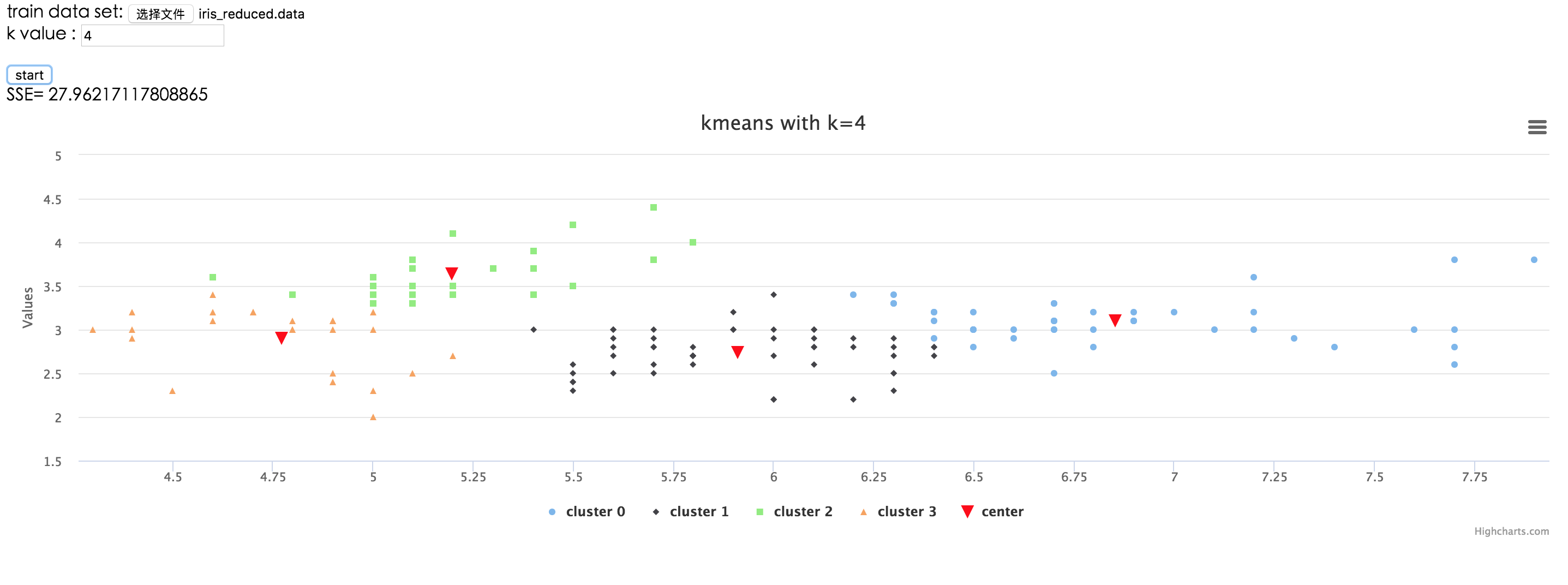
**Assignment 3**

潘瑞峰 1641478

**How to Run**

1. Open ‘kmeans.html’ with browser
2. Click ‘choose file’ button to choose one of the two datasets (iris\_reduced.data and wine\_reduced.data)
3. Input the k value (default 4)
4. Click the ‘start’ button
5. The result will show as the following picture.

**Data Sets**

1. Name: Iris Data Set

Attribute:

**a**) sepal length in cm.

**b**) sepal width in cm.

**c**) petal length in cm.

**d**) petal width in cm

**e)** class

1. Name: Wine Data Set

Attribute:

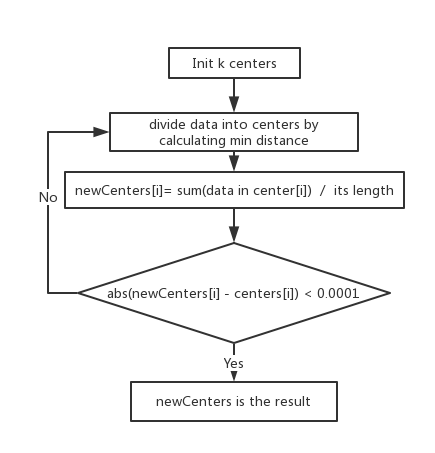
1. Alcohol
2. Malic acid
3. Ash
4. Alcalinity of ash
5. Magnesium
6. Total phenols
7. Flavanoids
8. Nonflavanoid phenols
9. Proanthocyanins
10. Color intensity
11. Hue
12. OD280/OD315 of diluted wines
13. Proline

**Data Preprocessing**

To show the results on two dimensions, I choose two of the attributes each data set.

I choose ‘sepal length in cm’ and ‘sepal width in cm’ in Iris Data Set; choose ‘Malic acid’ and ‘Ash’ in Wine Data Set.

**Design of the Program**



(Picture is drawn on processon and its url:

<https://www.processon.com/view/link/58538b35e4b097c6e69c1f0a>)

1. Choose first k points as the center
2. For data[i], if data[i] is nearest to center[j], then data[i] belongs to center[j]
3. Calculate new center[j] = sum of data in center[j] / its length
4. If change of old and new center is large than 0.0001, repeat step 2 and 3

**Result**

Result will be shown on the web, when you put your mouse on one point, it will show the coordinate. You can see how well the result is directly.

I also show the SSE of the result, you can change the k value to adjust the SSE.

**Limitations and improvements**

The first k points are chosen as the initial centers. To improve it, I should use different initial centers and choose the best ones whose SSE of result is minimal.

**Files**

1. iris\_reduced.data: dataset after Data Preprocessing
2. wine\_reduced.data: dataset after Data Preprocessing
3. kmeans.html: user interface of the program
4. kmeans.js: kernel of the program
5. origin\_data: origin dataset
6. lib: library of high-charts which is used to draw picture
7. reports.pdf: reports