Clustering

K-means Clustering

The following document shows k-means clustering of the'iris' data. At first, we remove species from the data to cluster. After that, we apply function kmeans() to iris2, and store the clustering result in kmeans.result. The cluster number is set to 3 in the code below.

```
> iris2<-iris
> iris2$Species<-NULL
```

```
> (kmeans.result<-kmeans(iris2, 3))
K-means clustering with 3 clusters of sizes 38, 62, 50

Cluster means:
  Sepal.Length Sepal.Width Petal.Length Petal.Width
1     6.850000    3.073684     5.742105    2.071053
2     5.901613    2.748387     4.393548    1.433871
3     5.006000    3.428000     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [38] 3 3 3 3 3 3 3 3 3 3 3 3 3 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2
 [75] 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 1 1 1 1 2 1 1 1 1
[112] 1 1 2 2 1 1 1 1 2 1 2 1 2 1 1 2 2 1 1 1 1 2 1 1 1 1 2 1 1 1 2 1 1 1 2 1
[149] 1 2

Within cluster sum of squares by cluster:
[1] 23.87947 39.82097 15.15100
 (between_SS / total_SS =  88.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss"
[6] "betweenss"    "size"
> |
```

The clustering result is then compared with the class label (species) to check whether similar objects are grouped together.

```
> table(iris$Species, kmeans.result$cluster)

              1  2  3
  setosa      0  0 50
  versicolor  2 48  0
  virginica  36 14  0
> |
```

The above result shows that cluster "setosa" can be easily separated from the other clusters, and that clusters "versicolor" and "virginica" are to a small degree overlapped with each other.

Next, the clusters and their centres are plotted.

Note that there are four dimensions in the data and that only the first two dimensions are used to draw the plot below.

Some black points close to the green centre (asterisk) are actually closer to the black center in the four dimensional space. We also need to be aware that the results of k-means clustering may vary from run to run, due to random selection of initial cluster centers.

```
> library(tm)
Warning message:
package 'tm' was built under R version 2.15.3
> library(SnowballC)
Warning message:
package 'SnowballC' was built under R version 2.15.3
>
```

```
> plot(iris2[c("Sepal.Length", "Sepal.Width")],col=kmeans.result$cluster)
> #plot cluster centres
> points(kmeans.result$centres[ ,c("Sepal.Length", "Sepal.Width")], col=1:3,
+ pch=8, cex=2)
```

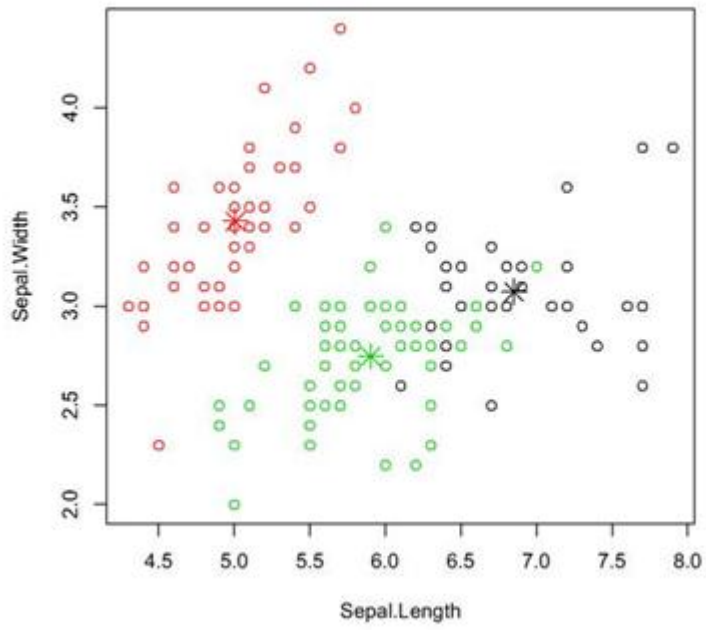

Fig: Results of K-Means Clustering

Text Clustering using K-means clustering: