CS-545-HW5:K-Means Clustering

Prerna Das

March 8, 2016

K-Means clustering algorithm was used to cluster and classify the OptiDigits data, originally from the UCI ML repository.

Experiment 1

K-means clustering was implemented on the training data set, so as to classify the data into 10 clusters. Iteration was stopped when all the cluster centers stoped changing. This was repeated 5 times and the run for which the lowest value of sum-squared error was obtained was used to classify the test data set.

```
cluster_class_10
```

```
##       cluster_name_10
##  [1,] "Clus1"         "5"
##  [2,] "Clus2"         "9"
##  [3,] "Clus3"         "6"
##  [4,] "Clus4"         "0"
##  [5,] "Clus5"         "7"
##  [6,] "Clus6"         "1"
##  [7,] "Clus7"         "3"
##  [8,] "Clus8"         "2"
##  [9,] "Clus9"         "8"
## [10,] "Clus10"        "4"
```

```
SSE_10
```

```
## [1] 2488472
```

```
SSS_10
```

```
## [1] 56710.77
```

```
mean_entropy_clustering_10
```

```
## [1] 0.8463589
```

```
confMat_test_10$table
```

```
##           Reference
## Prediction   0   1   2   3   4   5   6   7   8   9
##          0 176   0   0   0   2   0   0   0   0   0
##          1   0  58  21   2   0   0   4   0  97   0
##          2   1   2 149   6   0   0   0   3  13   3
##          3   0   0   1 151   0   1   0   8   7  15
##          4   0   5   0   0 162   0   0   6   8   0
##          5   0   0   0   0   1 144   1   0   0  36
```

```
##          6   1   0   0   0   1   0 176   0   3   0
##          7   0   5   0   0   1   4   0 166   3   0
##          8   0   8   1   5   0   4   2   2 121  31
##          9   0  23   0   5   0   6   0   5   1 140
```

```
confMat_test_10$overall
```

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##      0.8030050      0.7811667      0.7838493      0.8211660      0.1407902
## AccuracyPValue  McnemarPValue
##      0.0000000            NaN
```

Summary Experiment 1

We see that the 10 clusters represent one of the class instance from 0 to 9.

The **SSE** for the K-means algorithm was 2488472, **SSS** was 56710.77 and the **mean entropy of clustering** was 0.846

The **accuracy on the training data** was 0.798 and the **accuracy** on the test data was 0.80

Experiment 2

K-means clustering with 30 clusters.

```
cluster_class_30
```

```
##       cluster_name_30
##  [1,] "Clus1"         "1"
##  [2,] "Clus2"         "7"
##  [3,] "Clus3"         "2"
##  [4,] "Clus4"         "6"
##  [5,] "Clus5"         "1"
##  [6,] "Clus6"         "7"
##  [7,] "Clus7"         "6"
##  [8,] "Clus8"         "9"
##  [9,] "Clus9"         "5"
## [10,] "Clus10"        "7"
## [11,] "Clus11"        "9"
## [12,] "Clus12"        "1"
## [13,] "Clus13"        "1"
## [14,] "Clus14"        "3"
## [15,] "Clus15"        "1"
## [16,] "Clus16"        "2"
## [17,] "Clus17"        "8"
## [18,] "Clus18"        "0"
## [19,] "Clus19"        "9"
## [20,] "Clus20"        "2"
## [21,] "Clus21"        "4"
## [22,] "Clus22"        "4"
## [23,] "Clus23"        "0"
## [24,] "Clus24"        "4"
## [25,] "Clus25"        "4"
## [26,] "Clus26"        "3"
## [27,] "Clus27"        "5"
```

```
## [28,] "Clus28"        "8"
## [29,] "Clus29"        "0"
## [30,] "Clus30"        "3"
```

```
SSE_30
```

```
## [1] 1816074
```

```
SSS_30
```

```
## [1] 675415
```

```
mean_entropy_clustering_30
```

```
## [1] 0.3510055
```

```
confMat_test_30$table
```

```
##           Reference
## Prediction   0   1   2   3   4   5   6   7   8   9
##          0 177   0   0   0   1   0   0   0   0   0
##          1   0 177   0   0   0   0   0   0   2   3
##          2   0   4 161   1   0   0   0   3   8   0
##          3   0   1   2 165   0   1   0   5   4   5
##          4   0   5   0   0 174   0   0   2   0   0
##          5   1   0   0   0   1 170   1   0   0   9
##          6   2   2   0   0   1   2 172   0   2   0
##          7   0   0   0   0   1   0   0 168   1   9
##          8   0  18   0   2   1   1   1   2 140   9
##          9   1   1   0   5   3   2   0   4   1 163
```

```
confMat_test_30$overall
```

```
##       Accuracy          Kappa  AccuracyLower  AccuracyUpper   AccuracyNull
##      0.9276572      0.9196136      0.9146891      0.9392070      0.1157485
## AccuracyPValue  McnemarPValue
##      0.0000000            NaN
```

Summary Experiment 2

We see that the 30 clusters represent one of the class instance from 0 to 9.

The **SSE** for the K-means algorithm was 1816074, **SSS** was 675415 and the **mean entropy of clustering** was 0.351.

The **accuracy** on the training data and the test data was 0.932 and 0.927 respectively

Discussion

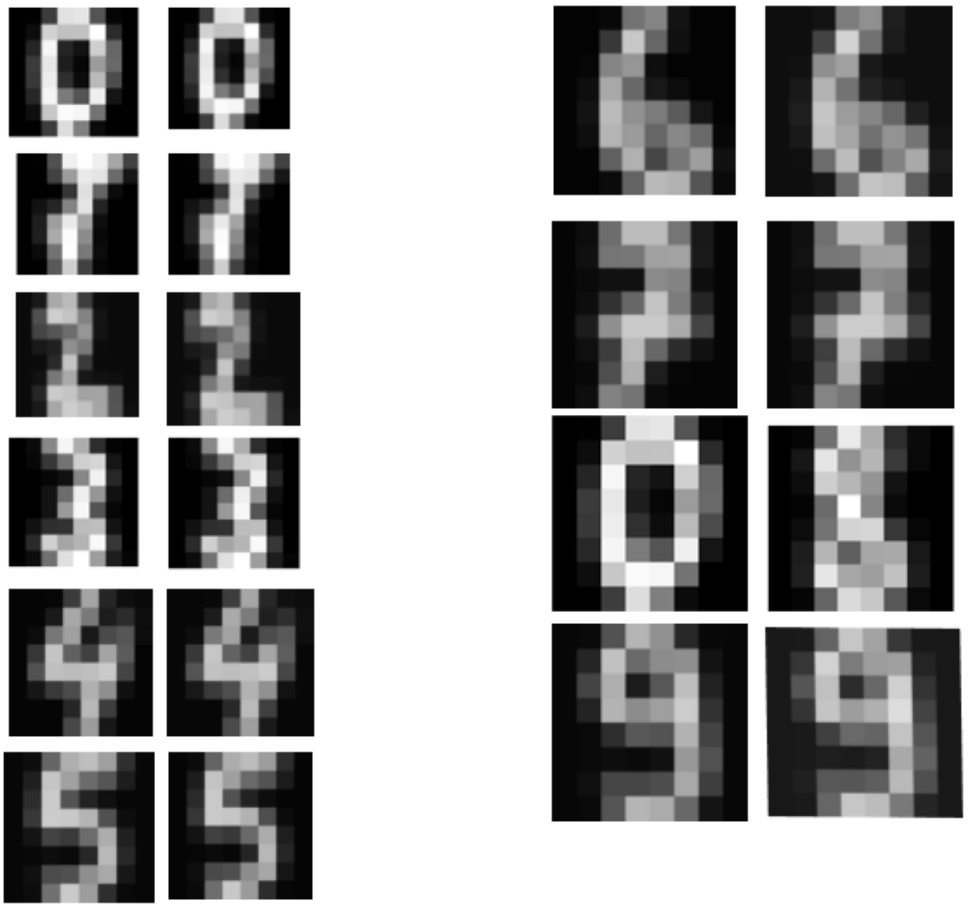
The above two experiments beautifully demonstrate the effect of increasing the number of clusters on SSE, SSS, mean entropy of clustering and accuracy achieved on the test data set.

When the number of clusters increases, the Sum Squared Error (SSE) decreases, the Sum Squared Separation (SSS) increases, and the Mean Entropy decreases, suggesting increasingly

uniform/pure clusters with increasing distance between the cluster centers. Accuracy also increases on increasing the number of clusters from 10 to 30.

Grayscale representation of the cluster centeroids features

For each image pair, the image on the left is from the cluster center obtained with K-means (K=10), while the image on the right is from the cluster center obtained with K-means (K=30).



Most of the digits were recognizable. The cluster centers did look like their associated digits, except for digit 8 in the K-means (K = 10), which looked more like the digit 0.