**Project 2: Clustering using K-means (Data Mining)**

Name – Navtejinder Singh Brar UTA ID – 1001874259

Name – Vamsi Krishna Pentakota UTA ID – 1001758105

**Description:**

There are three datasets named “Hourly\_2006”, “Hourly\_2009” and “Hourly\_2010” on which the clustering is done. We are clustering on the month of March in the datasets described earlier. For the preprocessing, we first extracted the data of March from the datasets and then removed all the attributes which are of no use for clustering the dataset except STN, Temp, DewP, STP, WDSP and yearModa\_hr. After preprocessing the number of objects for 2006 changed from 2111112 to 180792, for 2009 from 1612776 to 134232 and for 2010 from 1649304 to 134016. We also computed the mean of the dataset both daily and monthly for 2006, 2009 and 2010.While computing the mean we replaced the missing values with “na” and then computed the mean based on four dimensions: Temp, DewP, STP and WDSP.

Table

Description automatically generated

Figure 1.

There were a lot of missing values in all the datasets, we first identified those missing values and then replaced them with the means that we computed given in the Figure 1. After that, we created a new dataset grouped on STN. As there was a huge difference in the values of all four dimensions (Temp, DewP, STP and WDSP) getting the correct clusters was not feasible we overcame that by normalizing the data. Then, we created a function called wss plot for both Euclidean and Pearson which plots the curve to identify the best k for clustering. In this wss function we used the default value of seed as 15 and maximum iterations to 25. Moreover, we also plotted the curves for both Euclidean and Pearson for all the datasets (2006, 2009 and 2010) . After that, we used the Kmeans function to cluster the data for both Euclidean and Pearson distance metrics and we clustered the data from k=2 to k=8 and seed for this clustering was set to 15. After forming the clusters, we plotted them on the graph using autoplot function. In addition to this, we added the clusters we got to the corresponding station number in the dataset (final\_2006, final\_2009 and final\_2010). After that, we imported the stations dataset and did some preprocessing on that by removing the repeating station values. Moreover, we created new datasets for both Euclidean and Pearson in which we merged the clustered dataset and station dataset and named them (final\_2006\_plot\_pearson and final\_2006\_plot\_euclid (changed the year no. for different years)). After merging the datasets we plotted the points on the Texas map using the ggmap library. We will discuss everything in detail as we move further in the report.

**Division of Labor:**

|  |  |  |
| --- | --- | --- |
| **Name** | **Topics covered** | **Time Spent** |
| **Navtejinder Singh Brar** | Pre-processing, mapping on Texas map, Clustering for both seed values, Analyzing | 25 hours |
| **Vamsi Krishna Pentakota** | Pre-processing, Jaccard, Clustering for seed value=15, Analyzing | 25 hours |

**1. Clustering for each year from k=2 to k=8 for both Euclidean and Pearson.**

**a. Euclidean (2006) seed =15**

For 2006 March we did clustering from k=2 to k=8 setting seed as 15 and found out that the Within SS decreases drastically from k=1 to k=2 and after that it remains constant so from this we can infer that the **best value of k (no. of clusters) for 2006 is 2**. We also plotted the clusters we got from k=2 to k=8(Shown in Figure 3.). Moreover, in Figure 3 we can notice that the data points for k=2 are skewed towards one cluster.

A picture containing chart

Description automatically generated

Figure 2.

Diagram

Description automatically generated

Figure 3.

**b. Pearson (2006) seed =15**

Again, we did the clustering on the same data using the Pearson distance metric with seed as 15 and got the same result for Within SS i.e. it dropped sharply from k=1 to k=2 and after that it remained constant for all other k values(Shown in figure 4.). So, for Pearson too we got the **best value of k as 2**. Shown in figure 4 are the clusters plotted on the graph using Pearson metric for values of k=2 to 8. Moreover, in Figure 5 we can notice that for k=2 the data points are almost equally distributed among two clusters.

A picture containing line chart

Description automatically generated

Figure 4.

Diagram

Description automatically generated

Figure 5.

**c. Euclidean (2009) seed =15**

For 2009 March similar pattern to 2006 was seen for Euclidean using the seed as 15. For 2009 too we identified the **best k as 2**(Shown in Figure 6). The clusters for value of k from 2 to 8 are shown in the Figure 7.

Chart, line chart

Description automatically generated

Figure 6

Diagram

Description automatically generated

Figure 7

**d. Pearson (2009) seed =15**

Similar curve to Euclidean (2009) was seen in terms of finding the best k using the Pearson metric. So, the **best value of k for Pearson (2009) is also** **2** (Shown in Figure 8). The clusters for value of k from 2 to 8 are shown in the Figure 9.

Chart, line chart

Description automatically generated

Figure 8

Diagram

Description automatically generated

Figure 9

**e. Euclidean (2010) seed =15**

For March 2010 the curve shown in figure 10 was plotted. After analyzing the graph, we discovered the **best k to be 2** . Moreover, we plotted the clustering we got for k=2 to 8 on graph and is shown in figure 11.

Chart, line chart

Description automatically generated

Figure 10(up) and Figure 11(down)

Diagram

Description automatically generated

**f. Pearson (2010) seed =15**

In case of Pearson we are getting the same value and graph of best k as 2010 (Euclidean) that is **2(best k)**. In addition, we have plotted the clustering from k=2 to 8 on the graph shown in Figure 13.##

Chart, line chart

Description automatically generated

Figure 12

Diagram

Description automatically generated

Figure 13

**2. Using the best k for each year and metric, changing the seed from 15 to 150**

So far what we have analyzed is that the best k for each year in our case is 2. Moreover, we are required to change the seed (we are changing it from 15 to 150) and cluster the data again. So, after doing that this is what we have analyzed.

**a. Euclidean (2006, 2009 and 2010)**

When we look at the figures 3,7 and 11 and compare them with Figure 14, we can notice that there is no change in clustering after changing the seed value from 15 to 150 for the best k that is 2.

Chart

Description automatically generated2006Chart, funnel chart

Description automatically generated2009A picture containing shape

Description automatically generated2010

Figure 14

**b. Pearson (2006, 2009 and 2010)**

When we look at the figures 5, 9 and 13 and compare them with Figure 15 we can see that there is no change in clustering after changing the seed value from 15 to 150 for the best k that is 2, except for the year 2009 in which we are observing a significant change in clustering. Moreover, the cluster size has also changed from 78 and 90 (when seed is 15) to 67 and 101(when seed is 150).

Chart

Description automatically generated2006Chart

Description automatically generated2009Chart

Description automatically generated2010

Figure 15

**3. For the best k compare the clustering choosing different centroids (Euclidean)**

**a. For 2006 (k=2)**

We are using two seed values (15 and 150) to cluster using Euclidean metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 16 and for seed value of 150 we are getting the different centroids as shown in Figure 17. The SSE for the year 2006 for both the seed values remained same i.e. 0.3862 (refer to Figure 22). Moreover, there is no difference in the clustering for both the seed values (refer to figures 3 and 14). And, the cluster size for both the seed values remained same that is 24 and 128.

**Text

Description automatically generatedText, letter

Description automatically generated**

Figure 16 Figure 17

**b. For 2009 (k=2)**

We are using two seed values (15 and 150) to cluster using Euclidean metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 18 and for seed value of 150 we are getting the different centroids as shown in Figure 19. The SSE for the year 2009 for seed value 15 is 0.6724 whereas, for seed value of 150 it is 0.4325 which is significantly lower (refer to Figure 22). Moreover, there is no difference in the clustering for both the seed values (refer to figures 7 and 14). And, the cluster size for seed value of 15 (2009) is 68 and 100 whereas, for seed value of 150 (2009) the cluster size is 62 and 106.

Text

Description automatically generatedText

Description automatically generated

Figure 18 Figure 19

**c. For 2010 (k=2)**

We are using two seed values (15 and 150) to cluster using Euclidean metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 20 and for seed value of 150 we are getting the different centroids as shown in Figure 21. The SSE for the year 2010 for both the seed values remained same i.e. 1.2774 (refer to Figure 22). Moreover, there is no difference in the clustering for both the seed values (refer to figures 11 and 14). And, the cluster size for both the seed values remained same that is 137 and 37.

Text

Description automatically generatedText

Description automatically generated

Figure 20 Figure 21

A picture containing line chart

Description automatically generated

Figure 22

**4. For the best k compare the clustering choosing different centroids (Pearson)**

**a. For 2006 (k=2)**

We are using two seed values (15 and 150) to cluster using Pearson metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 23 and for seed value of 150 we are getting the different centroids as shown in Figure 24. The SSE for the year 2006 for both the seed values remained same i.e. 0.3328 (refer to Figure 29). Moreover, there is no difference in the clustering for both the seed values (refer to figures 5 and 15). And, the cluster size for both the seed values remained same that is 83 and 69.

Text

Description automatically generatedText

Description automatically generated

Figure 23 Figure 24

**b. For 2009 (k=2)**

We are using two seed values (15 and 150) to cluster using Pearson metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 25 and for seed value of 150 we are getting the different centroids as shown in Figure 26. The SSE for the year 2009 for seed value 15 is 0.1077 whereas, for seed value of 150 it is 0.6054 which is significantly higher (refer to Figure 29). Moreover, there is a considerable difference in the clustering for both the seed values (refer to figures 9 and 15).And, the cluster size for seed value of 15 (2009) is 78 and 90 whereas, for seed value of 150 (2009) the cluster size is 67 and 101.

Text

Description automatically generatedText

Description automatically generated

Figure 25 Figure 26

**c. For 2010 (k=2)**

We are using two seed values (15 and 150) to cluster using Pearson metric. For the first seed value that is 15 we are getting the centroids as shown in Figure 27 and for seed value of 150 we are getting the different centroids as shown in Figure 28. The SSE for the year 2010 for both the seed values remained same i.e. 0.4987 (refer to Figure 29). Moreover, there is no difference in the clustering for both the seed values (refer to figures 13 and 15). And, the cluster size for both the seed values remained same that is 99 and 75.

Text

Description automatically generatedText

Description automatically generated

Figure 27 Figure 28

Chart, line chart

Description automatically generated

Figure 29

**7. Visualization of clusters on the Texas map.**

**a. For year 2006, March and with seed =15 for Pearson metric** **with k=2**

Two clusters can be seen in Figure 30 the first cluster is on the north and west region of the Texas and the second cluster is on the south and south-west region of Texas.

Map

Description automatically generated

Figure 30

**b. For year 2010, March and with seed =15 for Euclidean metric with k=2**

Two clusters can be seen in Figure 31. The first cluster is in north and north and west of the Texas and the second cluster is in south and east on the Texas map

Map

Description automatically generated

Figure 31

**8. Three problems encountered and solved.**

**a.** The data set for the year 2010 was different as compared to 2006 and 2009 in case of delimiters (ingesting into R). In 2010 dataset for headers the delimiter was whitespace but after that the delimiter was comma. We solved this by changing the structure of the data using python.

**b.**