NAME : OMAR HASAN MOHIUDDIN

ID: 1001752543

PROJECT 2 REPORT

RESULTS AND ANALYSES

I. Analysis of Clusters from different seeds using SSE

We compare the SSE and total SSE for clusters with with different starting points (seeds = 40 and 10). However, we do not compare the SSE across different metrics since the calculation of error is different for each metric. In general, the clusters with a lower total SSE are better since they have an overall better intra-cluster similarity

Year = 2008, Metric = Euclidean

Seed = 40

22.65115 155.13575 634.26916 730.17541 102.35677 277.37988 = 1921.968

Seed = 10

487.1645 337.1991 637.6026 254.9181 141.5044 127.5369 = 1985.926

Result: seed = 40 has the better total SSE as well as clusters with much lower individual SSE compared to seed =10

Year = 2008 Metric = Pearson

Seed = 40

3.367145e-11 2.739804e-10 4.216380e-10 9.258739e-12 2.969871e-10 4.153360e-10 = 1.450872e-09

Seed = 10

4.136414e-10 2.740856e-10 3.341244e-11 4.216414e-10 2.970384e-10 9.262824e-12

= 1.449082e-09

Result: seed = 10 gives the better total SSE although the difference in the total SSE for both sets of clusters are almost the same

Year = 2009 Euclidean

Seed = 40

137.2803 141.2410 400.2655 288.0357 329.7699 49.0550 = 1345.647

Seed = 10

215.54418 133.11254 331.86507 279.78498 64.12892 449.96856 = 1474.404

Result: seed = 40 gives a significantly better result for total SSE and the individual SSE for its clusters are comparatively lower. Both seeds for 2009 have lower SSE than the results of 2008.

Year = 2009 Pearson

Seed = 40

1.419230e-09 2.919674e-10 7.343584e-10 2.619242e-06 5.353054e-10 7.950224e-10

= 2.623018e-06

Seed = 10

3.402224e-06 7.376890e-10 5.349796e-10 1.415498e-09 2.940712e-10 7.926924e-10

= 3.405999e-06

Result: seed = 40 gives a better total SSE but both sets of clusters have a much higher SSE compared to the previous year due to a single cluster in seed =10 and seed = 40 having significantly larger SSE compared to the others.

Year = 2010 Euclidean

Seed = 40

921.1937 194.4178 306.2180 440.6570 615.8495 378.0248 = 2856.361

Seed = 10

378.0623 617.3860 440.3847 922.1196 194.4092 306.2542 = 2858.616

Result: The SSE for both seeds are almost the same as well as individual SSE of clusters from seed = 40 has a counterpart with identical SSE in seed 10.

Year = 2010 Pearson

Seed = 40

6.530353e-11 6.664737e-13 9.175011e-13 3.506226e-10 1.035049e-09 1.475262e-09

=2.927821e-09

Seed = 10

9.382168e-12 7.173209e-10 4.198075e-11 2.783165e-11 1.988897e-11 6.632773e-11

= 8.827322e-10

Result: Seed = 10 has a lower total SSE and is consistently lower across its individual clusters compared to seed = 40

II. Analysis of Clusters from different metrics using Jaccard Coefficient

The calculation of Jaccard coefficients for a set of clusters determines the similarity of the weather stations in given cluster with another set of clusters. The general understanding of the value of the Jaccard coefficient vector is, given two sets of clusters M1 and M2, the Jaccard coefficient vector of M1 with M2 gives the highest similarity of each cluster in M1 with all the clusters in M2. Since this calculation is for M1, the same calculation for M2 with M1 is repeated.

For seed = 40

"2008 Euclidean with Pearson"

0.6055684 0.5214903 0.7441365 0.9002357 0.4963528 0.7547974

"2008 Pearson with Euclidean"

0.9002357 0.9002357 0.8867925 0.9002357 0.8605174 0.8476866

Analysis: The similarity of the clusters in Euclidean with Pearson ranges from 0.5 to .9 whereas the similarity of Pearson with Euclidean is consistently high. This could be due to the clusters for Pearson including a large variety of stations which match with a single cluster in Euclidean with a similar station distribution.

"2009 Euclidean with Pearson"

0.7168845 0.5071895 0.5016340 0.7686275 0.8884532 0.5094771

"2009 Pearson with Euclidean"

0.5631449 0.8884532 0.8884532 0.8753816 0.8250547 0.8884532

Analysis: 2009 shows a similar trend to 2008 with lower similarity for Pearson and a slightly higher similarity for Euclidean. A reason for this change can be that the sparsity of the station vectors for a cluster in Euclidean increased decreasing the maximum Jaccard coefficient for Euclidean and decreasing the overall Jaccard coefficients for Pearson.

"2010 Euclidean with Pearson"

0.7441365 0.5163148 0.6416788 0.7234878 0.9002357 0.4966895

"2010 Pearson with Euclidean"

0.9002357 0.9002357 0.9002357 0.9002357 0.9002357 0.6613419

Analysis: 2010 shows an almost similar trend to 2008 with Pearson having higher similarity with clusters in Euclidean. The reason for this could be that the sparsity of the station vectors for each cluster in Pearson has decreased whereas little change has been made to Euclidean.

Analysis for seed = 10: Since the vectors for each year and metric are almost identical for seed=40, the analysis for each year in seed = 40 applies to the corresponding year in seed=10

Below are the results for seed = 10

"2008 Euclidean with Pearson "

0.9002357 0.7996858 0.6941982 0.7336999 0.4980361 0.5214903

"2008 Pearson with Euclidean "

0.8476866 0.9002357 0.9002357 0.8867925 0.8605174 0.9002357

"2009 Euclidean with Pearson "

0.4980392 0.7369281 0.8884532 0.7686275 0.5094771 0.5643791

"2009 Pearson with Euclidean "

0.8753816 0.8884532 0.8250547 0.5631449 0.8884532 0.8884532

"2010 Euclidean with Pearson "

0.4966895 0.9002357 0.7234878 0.7441365 0.5074627 0.6416788

"2010 Pearson with Euclidean "

0.9002357 0.9002357 0.9002357 0.9002357 0.9002357 0.9002357

III. Year-wise analysis of Jaccard Coefficients

For Seed = 40 (since this has shown better clustering results)

"2008 TO 2009 EUCLIDEAN"

0.6950454 0.7428246 0.7986522 0.9834350 0.7570057 0.9805584

"2009 TO 2010 EUCLIDEAN"

0.9789112 0.6942714 0.7494258 0.9600312 1.0000000 1.0000000

Analysis:

2008 to 2009: 2 of six cluster of stations have almost identical weather in the next year . One of the clusters (jaccard coeff = 0.695) has noticeable change in stations and the remaining have high similarity with the next year.

2009 to 2010: Majority of the weather trends remains unchanged showing consistent weather patterns for those stations in September. Two sets of stations see a slight change in weather pattern across the year.

Between 2008 and 2010: The weather for 4 of 6 clusters (4 sets of stations) follows almost identical trends across these years whereas 2 sets of stations show a slightly more erratic pattern in weather.

"2008 TO 2009 PEARSON"

1.0000000 1.0000000 0.9850746 1.0000000 0.9558972 0.9416452

"2009 TO 2010 PEARSON"

0.6359994 1.0000000 1.0000000 0.9846154 0.9254621 1.0000000

Analysis:

2008 to 2009: Almost identical weather for each station with a few stations having slightly different weather.

2009 to 2010: One set of stations show a large change in the weather pattern across this year while the other stations experience weather similar to the previous year

Between 2008 and 2010: The weather for most stations remains largely the same however from 2009 to 2010 the weather for a certain set of stations changes noticeably (also observed the same with Euclidean metric)

•Overall Status. After preprocessing of data, I computed the year-wise calculations in a single generic function which gave outputs for the plot of points for that year on a Texas map, the station vector for each cluster used to calculate the Jaccard coefficient and cluster data for each cluster.

•File Descriptions.

kmeans.R : used to compute and plot the required data in the project description

Results.txt: Raw output of data (since running kmeans.R takes a long time)

Pearson year-wise.jpg and Euclidean year-wise.jpg: merged plots of stations belonging to same clusters for each year and seed.

•Division of Labor. Single member team so no division of labor. About 30 hours spent on the project in total.

•Problems encountered:

1. In the preprocessing of data, the dataset had missing column names. Column names had to be manually filled into the dataset.
2. Inconsistent separators in the 2010 dataset which were a combination of wither commas or just whitespaces. Replaced all commas in the dataset with whitespaces
3. Repetition of kmeans for each year became tedious. Created generic functions to calculate data and plots common to each year