CS60050 MACHINE LEARNING

Here I present a short report cum readme for the project

MINI PROJECT -3 REPORT

[Mini Project 3] **Project Code : CS4**

Cricket Format Clustering using Single Linkage Hierarchical Clustering Technique

Project Duration: 08-Mar-2021 ~~ 22-Mar-2021
Submission Information: (via) CSE-Moodle

Project Code – CS4
Project Title - Cricket Format Clustering using Single Linkage Hierarchical Clustering Technique
Submitted by – Deepak Mewada
Roll Number – 20CS91P02

Note : Here we have discussed results brief and shown corresponding outputs. For more details kindly refer to our program file (.ipynb).

Configuration of system and time taken to run

Overall time taken to run this program on a system with the following configuration is **63.67 seconds**.

The system configuration are as follows:

• **OS:** Windows 10 Home (64-Bit)
• **Processor:** Intel i3 @ 2.20 GHz
• **RAM:** 4 GB

Details to Read the Documents Submitted

- **20CS91P02_CS4.ipynb**: It is the main program code file that can be opened and run in a google colab environment. Please make sure that the source file *cricket_4_unlabelled.csv* is uploaded in the same temporary folder of colab. Otherwise, the program will generate a "file does not exist error".

  Once run, this program will generate the following files in the root folder where the .ipynbfile is located:

File name	Details
cricket_4_labelled_3.csv	CSV file with instance-wise information of clusters in case of 4 clusters for K-Means clustering algorithm.
cricket_4_labelled_4.csv	CSV file with instance-wise information of clusters in case of 4 clusters for K-Means clustering algorithm.
cricket_4_labelled_5.csv	CSV file with instance-wise information of clusters in case of 5 clusters for K-Means clustering algorithm.
cricket_4_labelled_6.csv	CSV file with instance-wise information of clusters in case of 6 clusters for K-Means clustering algorithm.
kmeans.txt:	Cluster-wise instances of the K-means in the format that has been asked for in the submission. This file is also a part of the ZIPPED submission folder.
agglomerative.txt:	Cluster-wise instances of the Agglomerative Hierarchical clustering algorithm with single linkage in the format that has been asked for in the submission. This file is also a part of the ZIPPED submission folder.

- **20CS91P02_CS4.pdf:** This is a run version of the .ipynb file, saved as PDF. This can be referred to for checking the outputs in case anyone does not prefer loading and running the .ipynb file. This includes all the theory, justifications asked for, code comments, etc. in Latexformat in markdown cells.

  Note: Since the initial clusters are randomly taken, the results of the K-Means clustering willdiffer minutely on each run.
- **kmeans.txt** and **agglomerative.txt**, as explained above.
- README_ProjectReport_CS4_20CS91P02.pdf: This file, which is a combination of both aREADME file as well as the Project Report with all observations and explanations.

Observations and Explanations

Optimal Value of K using Silhouette Coefficient

For K-Means algorithm, the various values of the Silhouette Coefficient for different values of K were observed as follows:

Value of K	Silhouette Coefficient using own model	Silhouette Coefficient using sklearn
3	0.312	0.309
4	0.309	0.302
5	0.294	0.236
6	0.2997	0.294

***Table 1**: Silhouette Scores for finding the Optimal Value of K*

We observe that the value of K for which the Silhouette Coefficient is maximum is 3 in both the cases (using own model as well as using sklearn), and hence, we take 3 as the optimal number of clusters going forward.

Table Details of Output files

Sr No.	File	Details
1	readme.txt	Readme file
2	kmeans.txt	Cluster-wise instances of the K-means in the format that has been asked for in the submission. This file is also a part of the ZIPPED submission folder.
3	agglomerative.txt	Cluster-wise instances of the Agglomerative Hierarchical clustering algorithm with single linkage in the format that has been asked for inthe submission. This file is also a part of the ZIPPED submission folder.

Note: click on text file to open it.

Jaccard Similarity

We observe that the clustering done by K-Means VS Agglomerative Hierarchical Clustering (AHC) gives similar results to an extent, but there are still a considerable number of instances thathave been differently clustered.

This can be proven using Jaccard Similarity (rounded off to 3 places of decimal), the matrix for which is given as under:

K-Means Cluster ID	AHC Cluster 0	AHC Cluster 1	AHC Cluster 2
K-Means Cluster 0	0.419	0.0	0.013
K-Means Cluster 1	0.136	0.454	0.084
K-Means Cluster 2	0.393	0.039	0.445

***Table 2**: Jaccard Similarity between K-Means and AHC Clusters*

Hence, we observe that the mapping of clusters with most common elements between them are as follows:

- K-Means Cluster 0 ↔ AHC Cluster 0
- K-Means Cluster 1 ↔ AHC Cluster 1
- K-Means Cluster 2 ↔ AHC Cluster 2

Interpretation of result:
Here we can understand from table according to the highest value of Jaccard coefficient we can conclude that:
a. k-means cluster number-0 is very similar to Agglomerative cluster number 0.
b. k-means cluster number-1 is very similar to Agglomerative cluster number 1
c. k-means cluster number-2 is very similar to Agglomerative cluster number 2.

You have reached the end of file.
Thank you!!!