CSCD94 Documentation

Tony Tan

June 2023

1 Application Explanation

This is an educational application designed to demonstrate the concept and application of Spectral Clustering, a powerful technique in the field of machine learning and data science. This technique is used to cluster data into relevant groups. Using our user-friendly interface, you can adjust parameters and observe their impact on the clustering process. For optimal visualization, we employ the make moons data-set from the sci-kit learn library. The adjustable include the sample size, the number of clusters, and the level of noise in the data. Users can experiment with various parameters to understand how these changes affect the results of the clustering process.

2 Theory description

Stanford Professor Andrew Ng's paper on spectral clustering offers deep insights into the mathematical and theoretical foundations of this technique. The paper thoroughly details the algorithmic process, which includes constructing a similarity graph, generating the Laplacian matrix, performing eigen decomposition, and finally, applying k-means clustering on the obtained eigenvectors to produce the clustered dataset. The paper underscores the ability of Spectral Clustering to manage non-convex clusters, a feature that elevates it above traditional methods like K-means for a broad array of applications.

3 Algorithm:

We define a set of points $S = s_1, ..., s_n$ and to partition them into k clusters.
1.Forming the affinity matrix
This is done by forming an affinity matrix A using formula:

$$A_{ij} = exp(-||s_i - s_j||^2/2\sigma^2).$$

The matrix A captures the euclidean distance between each point so we can measure how close each point is to others.

2.1Forming Diagonal(Degree) Matrix
The diagonal matrix is calculated using:

$$\forall i,\ D_{i,i} = \sum_{j=0}^{n-1} A_{i,j} \ \ and \ \ \forall j, j \neq i,\ D_{i,j} = 0$$

It is basically a degree matrix.

2.2Forming Laplacian Matrix
The Laplacian matrix captures the important property of the graph. It retains the information of connectivity between points and also the distance between points. It's calculated using the following:

$$L = D^{-1/2} A D^{-1/2}$$

3.The k largest eigenvalues and eigenvectors of the Laplacian Matrix
Depend on the number of clusters k we intend to have, we first find all the eigenvalues from the Laplacian matrix $x_1, ..., x_n$. And then we take the k largest eigenvalues from it, namely $y_1, ..., y_k$. Then from $y_1, ..., y_k$, we calculate the corresponding eigenvectors from it, namely $v_1, ..., v_k$. And we form the matrix $Y = [v_1, ..., v_k]$.

4.Normalize eigenvectors and apply K-means
We then form the normalized matrix N by performing normalization on the matrix Y above. We do this by making each row from Y to have unit length. Namely, $N_{i,j} = Y_{i,j}/(\sum_j Y_{i,j}^2)^{1/2}$. We treat each row from N as a point, and we apply K-means to N to obtain the labels by using K-means.

5.Cluster assignment
Lastly, we use the labels' information from K-means to assign each original point a cluster. Namely, we assign point s_i to cluster j if and only if K-means assigned row i of the matrix N to cluster j.

4 Application

4.1 Stack

The application is made using html, css, javascript with Flask backend.

4.2 Frontend

The front-end consists of homepage, step1 to step5 explanation and all-in-one explanation. The application also leverages the the extensive list of pre-designed components from Bootstrap version 4.5.2.

4.3 Back-end

The backend is done using Flask to handle *Post/GET* requests with libraries from sci-kit learn data-sets, numpy/scipy for matrix calculation and matplotlib for graph display.

5 Plan for Final version:

5.1 Better Algorithm Explanation

To perfect the theory-explanations in the application. I believe there are more background readings need to be completed in order to explain the algorithm in better/more intuitive way. Materials included revisiting the "On Spectral Clustering: Analysis and an algorithm", Fan R. K. Chung Graph Theory and some explanations from reliable sources, etc.

5.2 Advanced features

The plan is the extend more features of the application. Namely, allow more data sets for users to choose from. For example, the make circles data-set from sci-kit learn library, and classical diagonal lines. Also, to allow more parameters for users to tune on, e.g. the gamma(γ) value for the Affinity matrix.