Data mining assignments 4 : Spectral clustering

Benedith Mulongo

October 2019

1 Spectral clustering : the algorithm

Spectral clustering is clustering methods applied to graph data. It is done by computing properties of graphs such as the degree matrix, the number of outgoing edges from a node. The Laplacian matrix the difference between the degree matrix and adjacency matrix.

The article *On Spectral Clustering: Analysis and an algorithm* [1], proposes a new method for clustering graph by firstly computing the similarity matrix A using Gaussian kernel, then calculating the Laplacian matrix L of the similarity matrix, then the eigenvectors of the Laplacian L are used as the data for K-means algorithms.

The general algorithm is presented below :

Given a set of points $S = \{s_1, \dots, s_n\}$ in $\mathbb{R}^l$ that we want to cluster into k subsets:

1. Form the affinity matrix $A \in \mathbb{R}^{n \times n}$ defined by $A_{ij} = \exp(-\|s_i - s_j\|^2/2\sigma^2)$ if $i \neq j$, and $A_{ii} = 0$.
2. Define D to be the diagonal matrix whose (i,i)-element is the sum of A's i-th row, and construct the matrix $L = D^{-1/2}AD^{-1/2}$.
3. Find $x_1, x_2, \dots, x_k$, the k largest eigenvectors of L (chosen to be orthogonal to each other in the case of repeated eigenvalues), and form the matrix $X = [x_1 x_2 \dots x_k] \in \mathbb{R}^{n \times k}$ by stacking the eigenvectors in columns.
4. Form the matrix Y from X by renormalizing each of X's rows to have unit length (i.e. $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$).
5. Treating each row of Y as a point in $\mathbb{R}^k$, cluster them into k clusters via K-means or any other algorithm (that attempts to minimize distortion).
6. Finally, assign the original point s_i to cluster j if and only if row i of the matrix Y was assigned to cluster j.

Figure 1: Algorithm

2 Data

For testing and analyzing the behavior of the algorithm, we have used two different datasets. The first dataset is real and shows the realtionship between different researchers [1], the second dataset is synthetic.

- **Relationship among researchers :** The figure below shows the graph of the real data, we can observe that the best number of clusters is either 4 or 5. 4 is the most obvious.

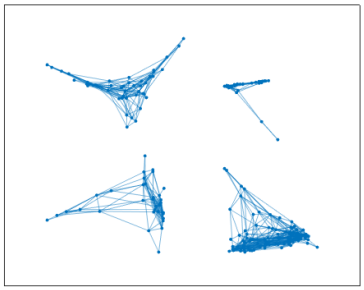

Figure 2: Real Data

- **Fictive data :** The figure below shows the graph of the synthetic data, we can observe that the best number of clusters 2.

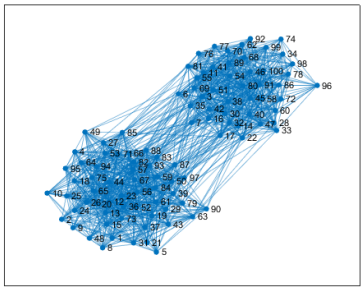

Figure 3: Fictive Data

[1] The dataset can be find at the website *http : //moreno.ss.uci.edu/data.htmlckm*

3 Observations

What we observe is that the result obtained from the similarity matrix using Gaussian kernel as described in the paper [1] is very poor. Using the adjacency matrix as the similarity matrix A, we obtain a more reliable result, that can be due to choose of the Sigma parameter.

Furthermore, it is noteworthy to observe that the K-means algorithms as implemented in Matlab are probabilistic such that the result obtained for one run may be different from a next run. Therefore is better to have a fixed center initialization by different methods, which is not done here.

4 Results

4.1 Real data

The number of K for K-means is 4. The graphs below show the result obtained.

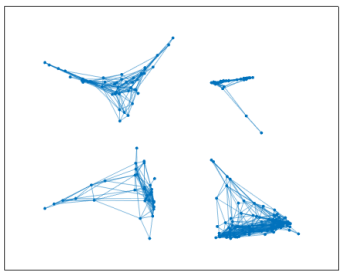

Figure 4: Real data graph

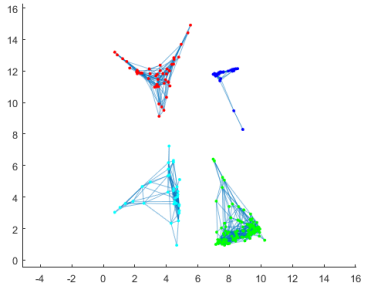

Figure 5: Graph of the clusters

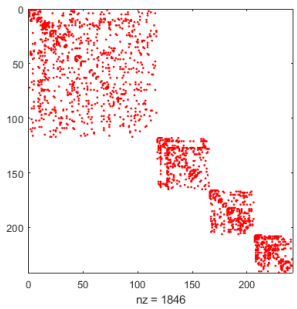

Figure 6: Sparsity Pattern

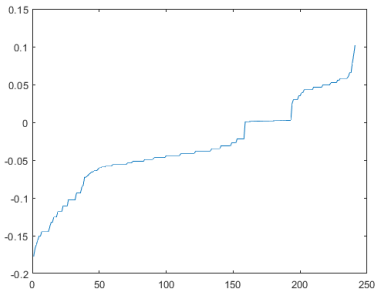

Figure 7: Sorted Fiedler Vector

4.2 Fictive data

For the fictive data, we have use K = 2. The result obtained is shown below.

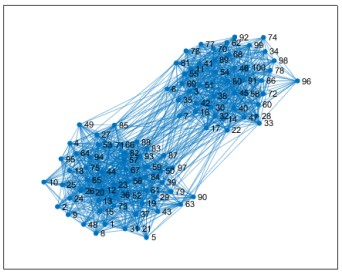

Figure 8: Result for TriestBase

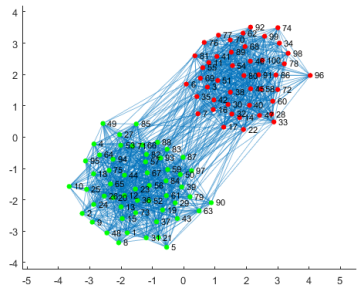

Figure 9: Result for TriestImpr

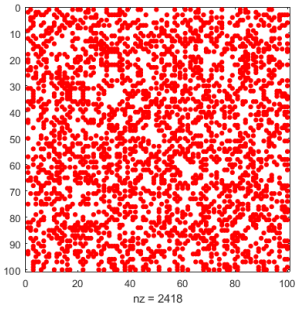

Figure 10: Sparsity Pattern

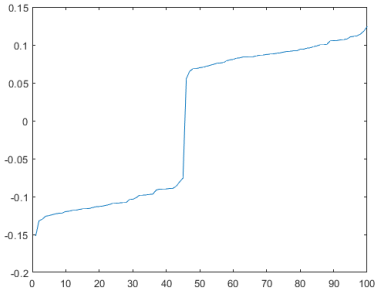

Figure 11: Sorted Fiedler Vector

References

[1] Y. Ng Andrew, I. Jordan Michael, and Weiss Yair. On Spectral Clustering: Analysis and an algorithm. *Encyclopedia of Machine Learning and Data Mining*, pages 1167–1167, 2017.