Clustering and Detection of Hub in a Complex Networks via the Laplacian Matrix

Choongrak Kim and Whasoo Bae *

Abstract— **In clustering (also known as unsupervised learning and class discovery), the classes are unknown a priori and need to be identified from the unsupervised data. The cluster analysis is concerned about estimating the number of classes and assigning each observation to a certain class. In this article we discuss a method for clustering via the Laplacian matrix. Also, based on a similar argument, we suggest a method for detecting hubs in a complex networks.**

Keywords: adjacency matrix, eigenvector, Fiedler vector, hub.

1 Introduction

Laplacian matrix is defined as subtracting the adjacency matrix from the degree matrix. The adjacency matrix is an $n \times n$ symmetric matrix in which the ijth component of it reveals the similarity of the connectivity between the ith and jth observation. On the other hand, the degree matrix is an $n \times n$ diagonal matrix where the ith diagonal element is the sum of the ith row of the adjacency matrix. The idea of using the Laplacian matrix in clustering is not new, and it can be found in, for example, Donath and Hoffman[1], Fiedler[2], Pothen, Simon, and Liou[3], Vishveshwara, Brinda, and Kannan[4], and Ding[5]. But, this approach is virtually unknown to statisticians and the suggested methods so far are not easily implemented for statistical data analysis. In this paper we suggest a useful algorithm via the Laplacian matrix for clustering. Also, we suggest a useful method for detecting hubs in a complex networks. Through illustrative examples based on real data sets we demonstrate the suggested algorithm is very effective.

Cluster analysis is a very important and most widely used tool for unsupervised type data. Good references on clustering are Everitt[6], Kaufman and Rousseeuw[7] and Gordon[8].

*Choongrak Kim is Professor, Department of Statistics, Pusan National University, Pusan, 609-735, KOREA (e-mail: crkim@pusan.ac.kr). Whasoo Bae is Professor, Department of Data Science, Inje University, Kyungnam, Korea, 621-749, KOREA (e-mail: wbae@stat.inje.ac.kr).

2 Clustering and Hub Detection

2.1 Laplacian Matrix

A graph $G = G(V, E)$ consists of a set of vertices V and a set of edges E. Two vertices v_i and v_j of a graph G are said to be adjacent if there an edge e_{ij} connecting v_i and v_j. A graph is called undirected(directed) if $e_{ij} = e_{ji}$ ($e_{ij} \neq e_{ji}$). The degree of a vertex v_i is defined as the number of adjacent vertices to v_i, and is denoted by deg_i. Graph (a) in Figure 1 has 5 vertices and 5 edges, and the degrees of vertices are 3,2,2,2,1, respectively.

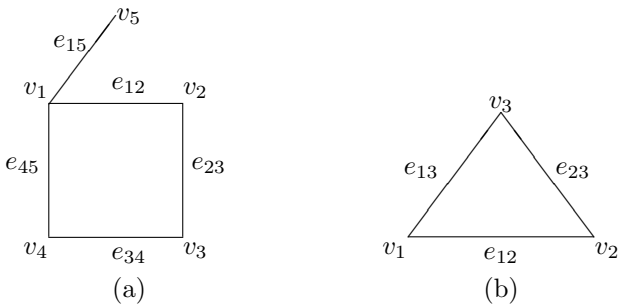

Fig. 1

The adjacency matrix $A = A(G)$ of an undirected graph G with n vertices is defined as an $n \times n$ symmetric matrix with components a_{ij}, where the diagonal elements a_{ii} are equal to zero for all $i = 1, 2, ..., n$. The adjacency matrix is called unweighted if $a_{ij} = 1$ for all $i \neq j$ for which edges e_{ij} is defined, and in general a_{ij} are not necessarily equal to 1. Graph (a) and (b) in Figure 2 have unweighted and weighted adjacency matrix, respectively. The Laplacian matrix of a graph G is defined as $L(G) = D(G) - A(G)$, where $D(G)$, called the degree matrix, is a diagonal matrix with the ith diagonal element $d_i = \sum_{j=1}^{n} a_{ij}$. Therefore, $D(G) = diag\{deg_1, ..., deg_n\}$ for unweighted adjacency matrix. Note that rank of the Laplacian matrix is $n - 1$. For detailed discussion on the graph theory and the Laplacian matrix, see, for example, Biggs[9], Deo[10], and Starng[11].

The Laplacian matrix has the following properties. It is symmetric and positive semidefinite, and therefore, it has n nonnegative eigenvalues. If the graph G is connected, then only one eigenvalue of them is 0 by the definition of

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

the Laplacian matrix, and all others are positive. If the graph is not connected, then the multiplicity of eigenvalue is equal the number of disconnected components. The eigenvector of L corresponding to the nonzero smallest eigenvalue is called the Fiedler vector in recognition of the pioneering works of Fiedler[2],[12].

2.2 Motivation on Clustering

Assume there are n objects or observations $\mathbf{x}_1, ..., \mathbf{x}_n$, where $\mathbf{x}_i$ is a $p-$vector. For example, in the gene expression matrix, $\mathbf{x}_i$ denotes expression of p genes for the ith patient. Also, $\mathbf{x}_i$ could be $p-$vector of (population, number of crimes per year, ..., average educational expenses per household) for the ith city. The goal is to partition n observations into an arbitrary number of groups such that observations in the same group have higher correlation or stronger connectivity than those in the other group. To use the Laplacian matrix in clustering, it is necessary to define the adjacency matrix. Let the adjacency matrix be some similarity measures for $\mathbf{x}_1, ..., \mathbf{x}_n$. For example, they could be Euclidean distance, Manhattan distance, and the Mahalanobis distance. In general the component of the adjacency matrix a_{ij} should reveal the closeness or degree of connectivity between $\mathbf{x}_i$ and $\mathbf{x}_j$.

Clustering can be achieved by minimizing the weighted sum of squares

$$Q = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij}$$

where $\mathbf{z} = (z_1, ..., z_n)'$ is unknown argument. To avoid the trivial solution $z_i = 0$ for all i, the constraint $\mathbf{z}'\mathbf{z} = 1$ is imposed. Also, the constraint $\mathbf{z}'\mathbf{1} = 0$, where $\mathbf{1} = (1, \cdots, 1)'$, is imposed since the minimum is invariant under translations. Therefore the problem can be rewritten as

$$\arg\min_{\mathbf{z}} \ \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i - z_j)^2 a_{ij}$$

subject to $\mathbf{z}'\mathbf{z} = 1$ and $\mathbf{z}'\mathbf{1} = 0$. To solve the problem, note that

$$\begin{aligned} Q &= \tfrac{1}{2}\textstyle\sum_{i=1}^{n}\sum_{j=1}^{n}(z_i^2 - 2z_i z_j + z_j^2)a_{ij} \\ &= \textstyle\sum_{i=1}^{n} z_i^2 a_{ii} - \sum_{j=1}^{n}\sum_{i\neq j}^{n} z_i z_j a_{ij} \\ &= \mathbf{z}'\mathbf{L}\mathbf{z} \end{aligned}$$

To minimize Q subject to $\mathbf{z}'\mathbf{z} = 1$, use Lagrangian method, i.e.,

$$\begin{aligned} T &= \mathbf{z}'\mathbf{L}\mathbf{z} - \lambda(\mathbf{z}'\mathbf{z} - 1) \\ \frac{\partial T}{\partial \mathbf{z}} &= 2\mathbf{L}\mathbf{z} - 2\lambda\mathbf{z} = \mathbf{0} \\ \Rightarrow & \quad (\mathbf{L} - \lambda\mathbf{I})\mathbf{z} = \mathbf{0} \end{aligned}$$

which yields a nontrivial solution $\mathbf{z}$ if and only if λ is an eigenvalue of $\mathbf{L}$ and $\mathbf{z}$ is the corresponding eigenvector. By multiplying $\mathbf{z}'$ on both sides, we have

$$\mathbf{z}'\mathbf{L}\mathbf{z} = \lambda$$

Therefore, the nonzero smallest eigenvalue and the associated eigenvector, which is called the Fiedler vector, yields the optimal solution.

On the other hand, to detect hub in a network, we need to maximize Q instead of minimizing Q in clustering. Therefore, the eigenvector corresponding to the largest eigenvalue contains the information on hubs, and we are only to find the components with large values (in absolute sense) in the eigenvector.

2.3 Hypothetical example

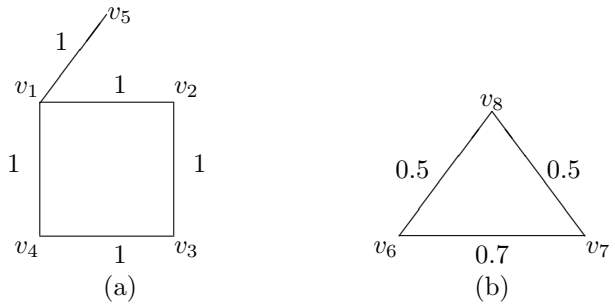

Fig. 2

Consider two graphs in Figure 2 where adjacencies between two points (1,2), (1,4), (1,5), (2,3), (3,4) are 1 in (a), and 0.5 in (6,8) and (7,8) and 0.7 in (6,7) in (b). Also, assume that all the adjacencies between $(v_1, \ldots, v_5)$ and (v_6, v_7, v_8) are 0.01, say. Then, the resulting eigenvalues and eigenvectors for the corresponding Laplacian matrix are given in Table 1. Since the rank of the Laplacian matrix is 7, one of 8 eigenvalues is zero. Note that the eigenvector corresponding to the nonzero smallest eigenvalue consists of two different values, 0.247 and -0.457. These values represent first 5 components and the next 3 components which match exactly to nearly separated graphs given in Figure 2. If we assume that all the adjacencies between $(v_1, \ldots, v_5)$ and (v_6, v_7, v_8) are zero, then the resulting Laplacian matrix has rank of 6. Therefore, two eigenvalues are zero and the corresponding eigenvectors are $(1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 1/\sqrt{5}, 0, 0, 0)$ and $(0, 0, 0, 0, 0, 1/\sqrt{3}, 1/\sqrt{3}, 1/\sqrt{3})$, respectively. Again, two graphs are perfectly separated by observing two eigenvectors corresponding to zero eigenvalues.

Also, we can detect hub from the eigenvector for the largest eigenvalue. Note that the eigenvector corresponding to the largest eigenvalue is

$$(0.702, -0.419, 0.338, -0.410, -0.202, 0, 0, 0)$$

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)

in Table 1. Hence, the 1st observation has the largest value 0.702, and it is hub in this network.

Table 1. Eigenvalues and eigenvectors for the Laplacian matrix of the combined graphs 2(a) and 2(b) in Figure 2.

Eigenvalues	0.00	0.08	0.90	$\cdots$	4.52
1	0.354	0.274	0.138	$\cdots$	0.702
2	0.354	0.274	-0.256	$\cdots$	-0.419
3	0.354	0.274	-0.438	$\cdots$	0.338
4	0.354	0.274	-0.256	$\cdots$	-0.419
5	0.354	0.274	0.812	$\cdots$	-0.202
6	0.354	-0.457	-0.000	$\cdots$	0.000
7	0.354	-0.457	0.000	$\cdots$	-0.000
8	0.354	-0.457	-0.000	$\cdots$	0.000

3 Conclusion and Future Works

The suggested method of clustering and detection of hub using the Laplacian matrix works well. In fact, the eigenvector corresponding to the non-zero smallest eigenvalue contains the information on clustering, and the eigenvector corresponding to the largest eigenvalue contains the information on hubs in a network.

To apply to the real data sets we need to refine the adjacency matrix by the hard-thresholding, say, and this area is worth pursuing as a future research.

References

[1] Donath, W.E. and Hoffman, A.J., "Lower bounds for partitioning of graphs," *IBM J. Res. Develop*, V17, pp. 420-425, 1973.

[2] Fiedler, M., "Algebraic connectivity of graphs," *Czech. Math. J.*, V25, pp. 619-633, 1973.

[3] Pothen, A. Simon, H.D., and Liou, K.P., "Partitioning sparse matrices with eigenvectors of graph," *SIAM J. Matrix Anal. Appl.*, V11, pp. 430-452, 1990.

[4] Vishveshwara, S., Brinda, K.V., and Kanna, N., "Protein structure: Insights from graph theory," *Journal of Theoretical and Computational Chemistry*, V1, pp. 187-211, 2002.

[5] Ding, C.H.Q., "Unsupervised feature selection via two-way ordering in gene expression analysis," *Bioinformatics*, V19, pp. 1259-1266, 2003.

[6] Everitt B., *Cluster Analysis,* Social Science Research Council by Heinemann Educational Books, 1980.

[7] Kaufman, L., Rousseeuw, P.J., *Finding Groups in Data: An Introduction to Cluster Analysis,* Wiley, 1990.

[8] Gordon S., *Controlling the State: Constitutionalism From Ancient Atherns To Today,* Cambridge: Harvard University Press, 1999.

[9] Biggs, N., *Algebraic Graph Theory*, Cambridge University Press, Cambridge, 1974.

[10] Deo, N., *Graph Theory with Applications to Engineering and Computer Science*, Prentice Hall, New Delhi, 1984.

[11] Strang, G.V., *Linear Algebra and its Applications,* Harcourt Brace Jovanovich, San Diego, 1974.

[12] Fiedler, M., "A property of eigenvectors of nonnegative symmetric matrices and its application to graph theory," *Czech. Math. J.*, V23, pp. 298-305, 1975.

ISBN: 978-988-19251-0-7
ISSN: 2078-0958 (Print); ISSN: 2078-0966 (Online)