Education

E-ISSN No : 2581-6195 | Volume : 1 | Issue : 1 | August 2018

A NEW SPECTRAL CLUSTERING APPROACH TO DETECTING COMMUNITIES IN GRAPHS

R. Moulay Taj[1], Z. Ait El Mouden[2], A. Jakimi[3], M. Hajar[4]

ABSTRACT

Recently, clustering is one of the most important approaches used for exploratory data analysis. It is one of the most widely used approaches for exploratory data analysis. Spectral Clustering (SC) is a technique which relies on the eigenstructure of a similarity matrix to partition points into disjoint clusters with the points in the same cluster having high similarity and the points in various clusters having low similarity.

Clustering nodes in graph is a general technique used in data mining for large network data sets. This paper presents an approach for the detection of communities from graphs with Spectral Clustering. In this research paper, for simulation and implementation, we used igraph package in language R.

KEYWORDS: Clustering, Spectral clustering, Similarity Matrix, Graph Laplacians, Language R

1. INTRODUCTION

In the last years, the data collected and processed know an exponential increase. Data Mining is a technique for extraction of knowledge in the massive data. Generally, the data mining methods are grouped into two families: the techniques of supervised learning and the technique of unsupervised learning. If we have a set of objects labeled and we know the possible classes we are talking about supervised learning. In the opposite case, it is unsupervised learning[1].

Clustering is one of the most unsupervised technique, concerned with the grouping of objects into clusters when their classes are not known in advance and the objects are not labeled [2, 3]. It's expected that objects of a cluster must have great similarity than the objects of others cluster. Basically to examine the similarity between objects the distance measurement is used.

In the literature, there are many clustering algorithms. The classical clustering algorithms are classified as hierarchical and partitional techniques[4]. the k-means algorithm considered the most popular algorithms that can be used[2]. Clustering algorithms have been used in different areas such as image processing, demographic study, crime detection, medicine and biology[4,5].

Spectral clustering has become popular modern clustering algorithms. It is mostly used for finding Communities in a graph (grouping nodes in a graph into clusters) from the similarity matrix[6,7, 8].

The paper is organized as follows: Section 2 describes the concept and process of spectral clustering. Section 3 presents different steps of our implementation. Section 4 presents the discussion of the results obtained. Finally, section 5 concludes the paper.

2. SECTRAL CLUSTERING

Spectral Clustering attracted more and more attention because of its sounds and good clustering resultants based on graph theory [9,7]. It's characterized by classification adapted to the search of the communities, even if it is not based on a probabilistic model but the spectral clustering has the advance to work on very large graphs[6].

Concerned with finding clusters in a set of graphs spectral clustering methods use the top eigenvectors of an affinity matrix, derived from similarities between data objects [2]. Each cluster delineated by its similarity which means that the objects.

The first step in spectral clustering is the create similarity graph with vertices are the data objects and edges are the affinities between data objects [10, 11].

This graph can be represented by an affinity matrix, where wij denotes the edge weight or affinity between vertices i and j. Creation of graph similarity is based on the concept of similarity sij>0 (inversely proportional to the distance).

The two mathematical objects used by spectral clustering are similarity graphs and graph Laplacians [6].

[1,4] Operational Research Team, Faculty of Science and Technology Errachidia, Morocco
[2,3] Software Engineering & Information Systems Engineering Team, Faculty of Science and Technology, Errachidia, Morocco

HOW TO CITE THIS ARTICLE:
R. Moulay Taj, Z. Ait El Mouden, A. Jakimi, M. Hajar(2018). A New Spectral Clustering Approach to Detecting Communities in Graphs, International Educational Journal of Science and Engineering (IEJSE), Vol: 1, Issue: 1, 01-05

Research Paper

2.1 Similarity graph

There are cases where data are not originally structured in a graph. In this case, a similarity graph can be constructed from these data.

We consider a classic data table of size n*p, i.e. n observations x1,x2,…..,xn with xi εR^p and it has a similarity measure between each pair of objects xi, xj. One of the most frequent similarity measures is given by the sigmoid function[9].For such, let d(i,j) be the dissimilarity between object I and j from the dataset, for example,the Euclidean distance.Then, the weight matrix W of a similarity graph G can be calculated by making :

$$wij = e^{-d(i,j)^2/\sigma^2}$$, if i≠j , and 0 otherwise.

The parameter σ has a high impact on the groups obtained. Different strategies have been investigated to find its best value. The primary goal of clustering is to divide the data objects into several groups such that objects in the same group are similar and objects in various groups are dissimilar to each other. The goal of constructing similarity graph is to model the local neighborhood relationships between the data objects. There are several popular constructions to transform a given set $x_1,...,x_n$ of data objects with pairwise similarities s_{ij} or pairwise distances dij into a graph. The several popular similarity graphs used in spectral clustering are the €-neighborhood graph, K-nearest neighbor graphs, and fully connected graph.

The €-neighborhood graph, here we set a scale ε >0 and we connect all points vi, vj where s_{ij}>= ε (whose distance value are smaller than ε) the graph constructed is an unweighted graph.

K-nearest neighbor graphs, where the idea is to connect each vertices to its k nearest neighbors. However, this yields a directed graph since the k-nearest neighbors relationship is not symmetric. If we want to construct an undirected KNN graph we can chose between the mutual KNN graph, where there is an edge between two vertices if both points are among the K nearest neighbors of the other one, and the symmetric KNN graph, where there is an edge between two vertices if one point is among the k nearest neighbors of the other one. In both cases edges constructed are weight by the similarity of their endpoints in order to form a directed graph.

The fully connected graph: Here we simply connect all vertices with positive similarity with eachother, and we weight all edges by sij. (The vertices vi ,vj are connected if i≠j and sij>0) as well the graph created is dense.

2.2 Graph laplacians

Laplacians matrices are the main objects for spectral clustering. In the following G is an undirected weighted graph,with weighted matrix W (size n*n), such that w_{ij}≥0. Let D is a diagonal matrix whose diagonal is (d1,.....,dn) with di is the valued degree of the node I in G i.e. di=$\sum_j$wij=$\sum_j$wji.

The two types of graph laplacians with their important properties are defined as the following[6]:

- The unormalized graph Laplacians

The unormalized graph Laplacian is defined as: L=D-W

Matrix L satisfies the following properties:

1. 1- For every vector uεR^n, we have
$$\mathbf{u^T L u} = \frac{1}{2}\sum_{i,j=1}^{n} \boldsymbol{Aij(ui - uj)^2}$$
2. L is symmetric and positive.
3. The smallest eigenvalue of L is 0, the corresponding eigenvector is the constant one vector 1.
4. L has n non-negative, real-valued eigenvalues 0 = λ1 ≤ λ2 ≤ .... ≤ λn.

The unnormalized graph Laplacian and its eigenvalues and eigenvectors can be used to describe many properties of graphs. One example which will be important for spectral clustering number of connected components and the spectrum of L. The multiplicity of the value own 0 is equal to the number of connected components of G. The Vectors indicators of these connected components are of the eigenvectors for the value own 0.

The study of the Laplacian spectrum of the graph thus makes it possible to determine simply the numbers of connected components of this graph.

The unormalized graph laplacian is defined as:

$$\boldsymbol{L_N = I - D^{-1/2} A D^{-1/2}}$$

3. APPROACH PROPOSED

The goal of our approach is the development and implementation of a data grouping process based on the spectral clustering algorithm and the igraph library [12]. The igraph library implements a good set of community detection algorithms, allowing researchers to easily apply them to data mining tasks[13]. This approach consists of five major steps; the first step named definition of data which can summarize by the collection of the data thus the description of these last, the second stage is the graphical representation of the data to be treated on the basis of the graph of similarity which one already mentioned. In the third step, we will present the different matrix representations of similarity graphs built in the previous part. The fourth step represents the application of the spectral clustering algorithm on the matrix obtained in the previous step in order to deduce the groups of our data. The fifth step is the interpretation of the results. In the following, we will describe these different stages with the examples.

3.1 Definition of data

The definition of data begins with the collection of all n individuals who make up our database. They can be noted: X1, X2 X3, ……..Xn. Individuals Xi may be heterogeneous. After the definition of n individuals, the next step is to define for each individual the whole of these components Xi = (x j)jε[1,p], a component xj of the given Xi is equivalent to an attribute in the relational model.

Example

We are interested in our simulation to a database of students of the Faculty of Sciences and Techniques of Errachidia where each student is described by a set of characteristics: CNE,

Name, Module stream, Section and the note. The student will be present by a vector with components.
Student = {CNE, Name, Module, stream, Note}

Attribute	Description
CNE	The national code of the student, in digital format
Name	The name of student, in text format
Module	The module study each module has its own code under forma XXX
Stream	The student stream has two possible values {BCG, MIP}
Note	Note of module, in numeric format

Table1: Description of attributes

3.2 Graphic representation

Spectral clustering algorithm consists of one significant step to construct a similarity matrix and the goal of constructing the similarity matrix is to model the local neighborhood relationships between the data vertexes. Consider we have a data set X = x1, ...,xN which we want to cluster into K clusters. To construct the affinity (or similarity) matrix S which measures the weights (or similarities) between all data points we can use the following formula:

$$s_{ij} = \mathbf{exp(- (xi-xj) /2\sigma^2)} \quad \text{If } i = j, \text{ else } s_{ii} = 0$$

From a similarity matrix we will create the graph of similarity defines in the following manner G = (V, E) where V is the set of nodes of the graph and E the set of edges between the nodes of V. There are three types of graphs of similarity (which they are already cited), the graphs of dense similarity (fully connected), the neighborhood de-neighborhood graphs (ε-neighborhood) and the graphs of the k nearest neighbors (k-nearest neighbor).

Example

Considering our student database which contains 20 observations, our goal is the classification of this observation. This classification requires in the first place a pretreatment to find the optimal set of relevant attributes. We will use the field CNE to label the nodes of the graph, then he will be neglected in the calculations of the distances as well as the attribute name because they bear no reversal important for the classification of nodes. To calculate the distance matrix of our example we will use the Euclidean distance for the numerical attributes namely note. For non-numeric attributes, the distance is 0 if the values are equal and 1 otherwise.

From the table of similarity, we can easily construct the graphs of similarity; the following figures represent the graph similarity:

- Figure 1: the graph of similarity dense, where all the edges will be present except in the cases of boucle.
- Figure 2: shows the construction of ε-neighborhood graphs with different scale values.
- Figure 3: the influence of the parameter K on the generated k-nearest neighbor graphs.

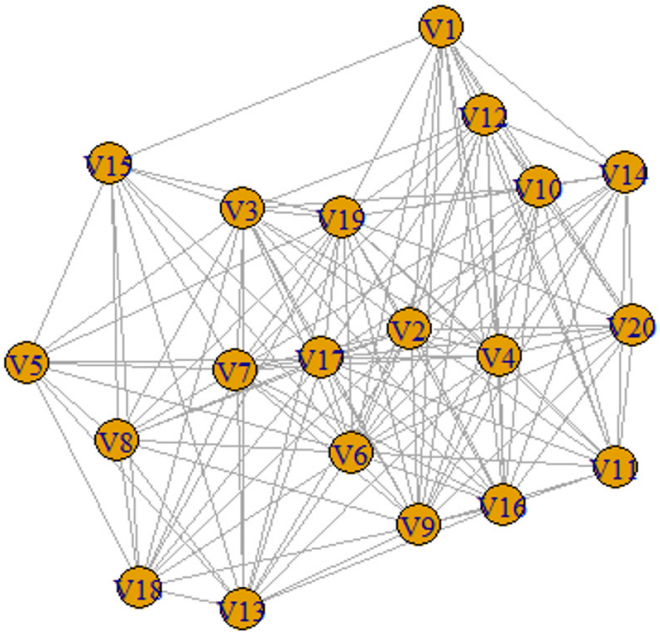
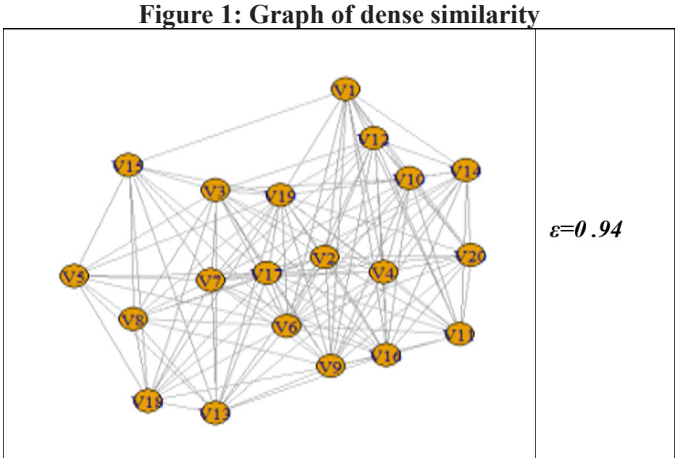


Figure 1: Graph of dense similarity



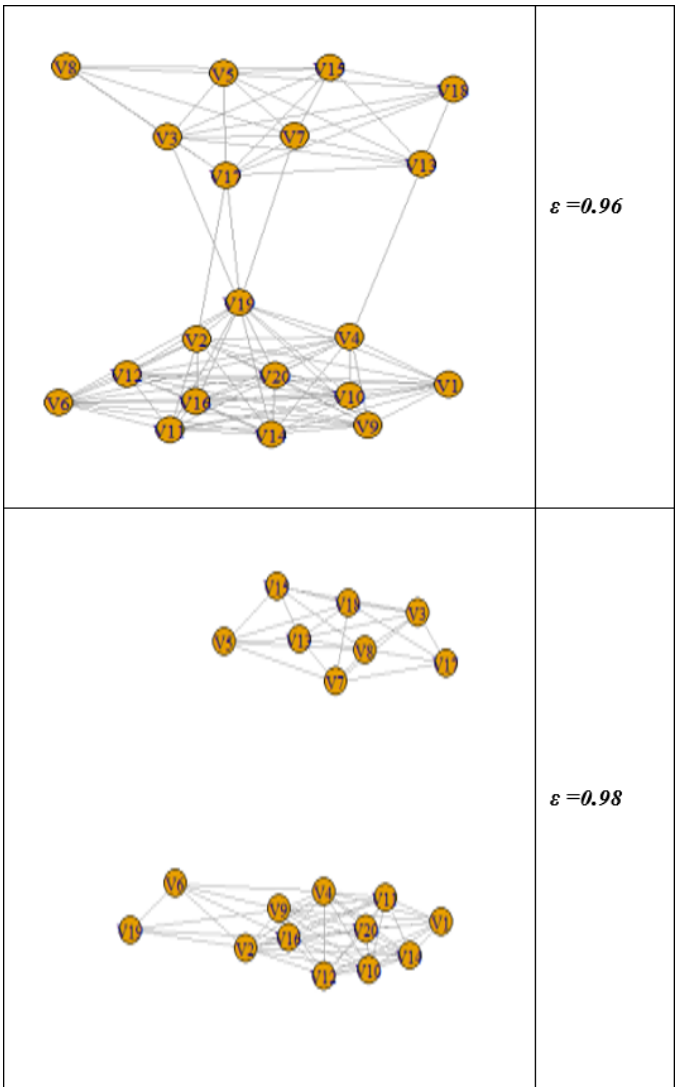


Figure 2: the construction of ε-neighborhood graphs with different scale values.

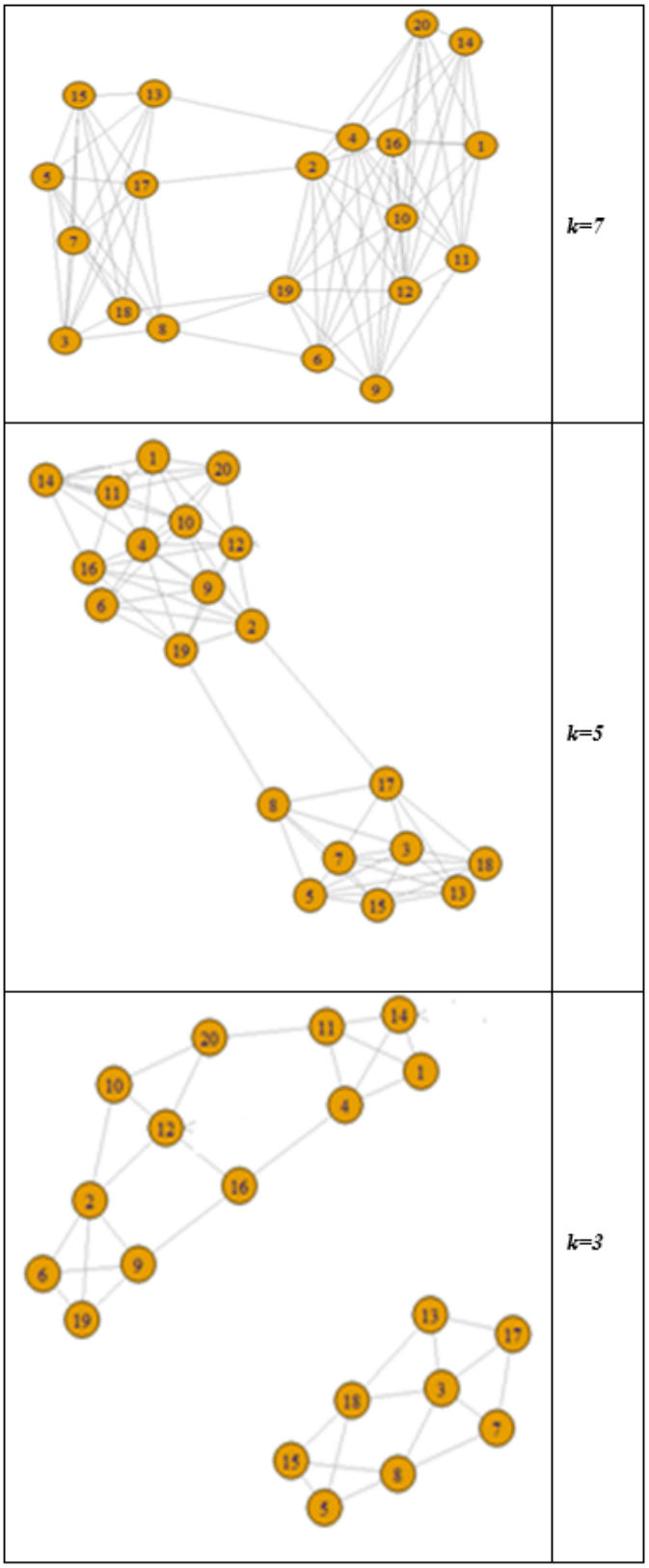


Figure 3: K-nearest neighbor graphs.

The limits of graph of similarity dense are numerous; the presence of all the edges is not an important information, in the case dealt with the presence of the edges with the weight almost zero adds no value to the graph result, on the contrary, it increases the time of generation of the graph as well as its complexity.

In the case of ε-neighborhood graphs, the ε parameter plays an important role in the quality of similarity graphs. The parameter K plays a crucial role in the results provided by the construction algorithm of k-nearest neighbor graphs. A higher value of K generated a higher number of links between the nodes, when a lower value of K risks of disappearing edges that carries information about the visualized data.

3.3 The matrix representation

Form the affinity matrix W we define D to be the diagonal matrix and construct the Laplacian matrix L, then obtains the eigenvectors and eigenvalues of L. Find x1, x2... xk, the k largest eigenvectors of L and form the matrix X= [x1, x2... xk] €Rnk by stacking the eigenvectors as columns. Construct the matrix Y from X by renormalizing each of X's rows to have unit length. We treating each row of Y as a point in Rk, cluster them into k clusters using any clustering algorithm. Finally assign the original point Si to cluster j if only if row i of the matrix Y has assigned to cluster.

3.4 Results interpretation

For the data classification of our database we will build the graph of similarity by ε-neighborhood graphs with scale values is 0.96 and the result of Spectral Clustering is the classification of the nodes in two clusters C1 and C2. C1 contains students from MIP stream and in C2 there are BCG stream students. The following figure represents the result of Spectral Clustering implement with igraph.

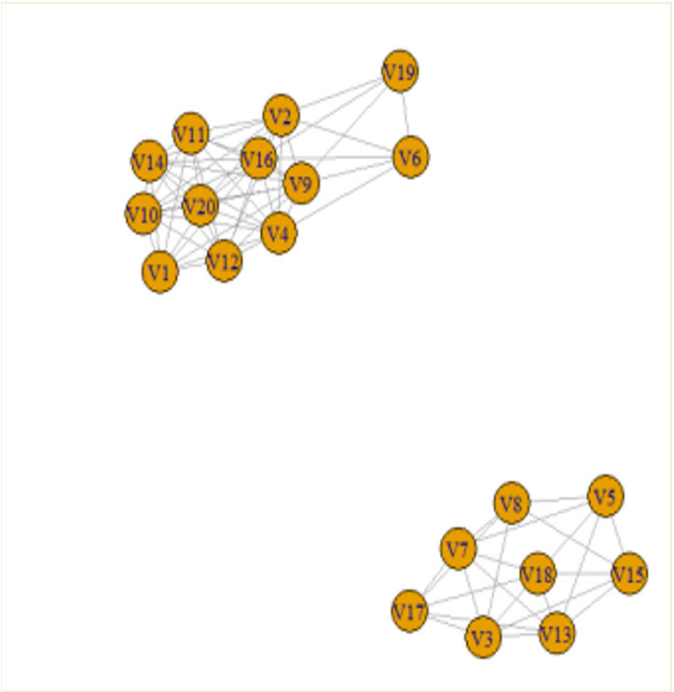


Figure 4: the result of Spectral Clustering implement with igraph.

The choice of the similarity graph and the selection of the parameters of the different phases of the processes play an important role in the result obtained. The modification of any parameter can generate different results.

4. CONCLUSIONS

In this paper, we presented clustering spectral as a method of classification of data modeled by graphs, in passing by the matrix representation of the graph of similarity and the spectral analysis of the matrices generated. We used a case study to see the behavior of algorithms from spectral clustering as well as the impact of the change of some parameters on the final results. Spectral Clustering not only serves the purpose of detecting communities in a network, it can also enable us to identify classes whose delimitations are non-convex as long as we are able to define a relevant measure of similarity (or affinity) between individuals.

5.REFERENCES

1. C. Drouin, "Datamining La famille des rois de France." http://jeanjacques.villemag.free.fr/
2. Ajit Kumar, Dharmender Kumar, S. K. Jarial, A Review on Artificial Bee Colony Algorithms and Their Applications to Data Clustering, Journal of Institute of Information and Communication Technologies of Bulgarian Academy of Sciences, V17, I3, 2017.
3. Studies in Classification, Data Analysis, and Knowledge Organization}. Springer, Heidelberg, 3–10. Bock, H.-H. , 2008.
4. A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," Inf. Sci. (Ny)., vol. 222, pp. 175–184, 2013.
5. C. D. Dan A. Simovici, Mathematical Tools for Data Mining. (Advanced Information and Knowledge Processing) , Book, Springer, 2008.
6. U. Von Luxburg, "A tutorial on spectral clustering," Stat. Comput., vol. 17, no. 4, pp. 395–416, 2007.
7. Z. Jingmao and S. Yanxia, "Review on spectral methods for clustering," Chinese Control Conf. CCC, vol. 2015–Septe, no. 20130093110011, pp. 3791–3796, 2015.
8. S. White and P. Smyth, "A Spectral Clustering Approach To Finding Communities in Graphs." pp. 76–84, 2005.
9. M. C. V Nascimento and A. C. P. L. F. De Carvalho, "Spectral methods for graph clustering – A survey," vol. 211, pp. 221–231, 2011.
10. S. Meenakshi and R. Renukadevi, "A Review on Spectral Clustering and its," pp. 14840–14845, 2016.
11. A. Y. Ng, M. I. Jordan, and Y. Weiss, "On Spectral Clustering: Analysis and an algorithm," Adv. Neural Inf. Process. Syst., pp. 849–856, 2001.
12. G. Csárdi and T. Nepusz, "The igraph software package for complex network research," InterJournal, Complex Syst., vol. 1695, no. 5, pp. 1–9, 2006.
13. F. B. De Sousa and S. P. Brazil, "Evaluating and comparing the igraph community detection algorithms," 2014, pp. 408–413.