Graphical Determination of Groups and Outliers in Distance-Based Cluster Analysis

Luis F. Rivera-Galicia
Department of Economics
University of Alcalá, Alcalá de Henares, SPAIN
e-mail: luisf.rivera@uah.es

Abstract

Cluster analysis is a popular unsupervised learning method. Its goal is to find a partition of a dataset of *N* objects into *k* well separated groups: elements within a group must be similar (in some sense) to one another, and different to elements in other groups. The fundamental problem of cluster analysis is to determine the real number of groups (*k*) in the dataset.
In this paper, a new method of clustering is presented, to simultaneously determine the number of groups and the clustering in a dataset. This method is based on graph theory. Dissimilarity data between objects is used to form a dissimilarity graph, in which the vertices are the objects in dataset, and the edges are weighted according to the dissimilarity between the objects. Two vertices are then connected by an edge, when the dissimilarity among them is under some certain threshold. A statistical procedure is proposed to determine the appropriate threshold to split the graph into its connected components. As an additional result, cases which are isolated can be considered as outliers and may need to be further analyzed. This method has been tested on some different datasets, and results obtained are analyzed taking into account the resulting clustering.

Keywords: clustering, number of groups, outliers.

1. Introduction

Cluster analysis is a popular unsupervised learning method which aims to find a partition of a dataset into *k* well separated groups (clusters). The whole collection of clusters in a dataset is known as clustering.
There are many ways in which clustering techniques can be classified (Gan, Ma and Wo, 2007). Clustering techniques can be hierarchical or partitional. Hierarchical clustering methods create a sequence of nested clusterings of a dataset, while partitional methods create a single clustering (flat clustering), based on distance between groups or on densities of groups. Clustering can be exclusive, overlapping or fuzzy. Exclusive clustering assigns each object of dataset to a single group, overlapping clustering assigns each object to some (maybe more than one) of the groups in dataset, and fuzzy clustering assigns to every object the probability of belonging to every group in dataset. Clustering can be complete or partial. Complete clustering assigns every object of dataset to a group, and partial clustering does not classify all objects in dataset (for details, see Rokach, 2010).
Choosing the number *k* of clusters is a general problem for all clustering algorithms. Indeed, the problem of determining the real number of clusters in a dataset is considered the "fundamental problem" of cluster analysis (Dubes, 1993, Hardy, 1996). As cluster analysis is an unsupervised learning method, there is no prior knowledge about the number of clusters in the dataset. This leads us to the problem of cluster validity. There are several measures to assess the quality of clustering, taking into account different criteria: internal, external or relative criteria.

In this paper, we are going to develop and analyze a clustering method based on graph theory methods, which will simultaneously determine the number *k* of groups and the resulting clustering. This procedure is partitional, exclusive and complete.
The rest of the paper is organized as follows: Section 2 contains the algorithm that will be used. Section 3 contains the application of the method to some datasets. Section 4 presents the discussion of the results and the main conclusions of the paper.

2. Method

2.1. Cluster analysis and graph theory

Let $O=\{O_1,O_2,...,O_N\}$ be a set of *N* objects, entities or patterns among which clusters are to be found. Each of these objects is a *p*-dimensional vector in a given feature space. Thus, we have a *Nxp* input data matrix *X*. In order to start the clustering procedure, a proximity matrix *P* has to be computed, containing the pairwise indices of proximity of a dataset. In what follows, we consider the proximity indices to represent a symmetric dissimilarity function between objects.

$$P=\left(d_{ij}\right)_{\substack{i\in\{1,...,n\}\\ j\in\{1,...,n\}}}.$$

Graph partitioning can be considered as data clustering using a graph model. Given the coordinates of the data objects and the dissimilarity between any two points, the symmetric matrix containing dissimilarities between all pairs of points (*P*) forms a weighted adjacency matrix of an undirected graph. Thus, the data clustering problem becomes a graph partition problem, through the search of graph connected components (Tan, Steinbach and Kumar, 2005).

2.2. The search for connected components

Let $G=(V,E)$ the similarity graph derived from dataset. Each vertex v_i in this graph ($v_i \in V$) represents a data object O_i. Two vertices are connected by an edge in *E* if the dissimilarity d_{ij} between the corresponding objects O_i and O_j is under a certain threshold, *d*.
Let $d_{\max}$ be the maximum dissimilarity index in the proximity matrix. The whole range of dissimilarities in the proximity matrix can be analyzed to get the best clustering in dataset (this range is $[0,d_{\max}]$, since 0 is in the proximity matrix, as it represents the dissimilarity of an object to itself).
Given a certain value of *d* ($d\in[0,d_{\max}]$), the *Adjacency matrix* of the graph at level *d* is the matrix $A^d=(a_{ij}^d)_{i,j\in\{1,...,n\}}$, whose elements are defined as follows:

$$a_{ij}^d=\begin{cases}1 & \text{if } d_{ij}<d\\ 0 & \text{otherwise}\end{cases},\ i,j\in\{1,...,n\}$$

The *degree of a vertex* $v_i \in V$ at level *d* is defined as the number of vertices in *V* that are connected to v_i at a lower dissimilarity level than *d*, that is, $degree(v_i^d)=\sum_{j=i}^{n} a_{ij}^d$ (note that every vertex v_i is connected with itself at every dissimilarity level $d>0$).
The *degree matrix* D^d is the diagonal matrix with the degrees of all nodes on the diagonal, considered at the dissimilarity level, *d*.
For each dissimilarity level *d*, the unnormalized graph Laplacian matrix, L^d, is defined as:

$$L^d=D^d-A^d.$$

Laplacian matrices are the main tools for spectral clustering associated to graph

partitioning. There exists a whole field devoted to the study of these matrices and their properties, called spectral graph theory.
The Laplacian matrix has the following properties (von Luxburg, 2007):

1. L is symmetric and positive semi-definite (thus its eigenvalues are non-negative).
2. L is a conservative matrix (0 is always an eigenvalue of L, the smallest, whose eigenvector is the *n-dimensional* 1 vector).

Fiedler (1973) proved that the multiplicity k of the eigenvalue 0 of L is equal to the number of connected components in the graph. At every value of d, the number of connected components in the graph can be computed through the number of null eigenvalues.
The procedure proposed in this work is to find the number of clusters that can be obtained at every dissimilarity level, k^d, and to analyze how the clustering fits the dissimilarity structure of data, to get the best clustering.

2.3. Clustering validation

The evaluation of clustering validity can be analized from three different points of view. Supervised measures, sometimes called external indices, refer to some external structure, and they need information not present in dataset. Relative measures are used to compare different clusterings of the same dataset. Unsupervised measures, or internal indices, use only information present in dataset.
We take this last approach, because it is useful only for partitional sets of clusters and its use in our case is very intuitive, as it can be interpreted as a correlation coefficient.
For every clustering (obtained depending on the varying value of d in the proximity matrix), we compute the Hubert's modified statistics, which is defined as the inverse of the point serial correlation coefficient between dissimilarity matrix and Adjacency matrix.

$$\Gamma = -corr\left(P, A^d\right).$$

This coefficient Γ belongs to the interval [0,1] and can be calculated only for two or more clusters. The larger the value of Γ, the better the clustering.

2.4. The algorithm

As a result of previous sections, the proposed clustering algorithm is the following:

Input: $X = \left(x_{ij}\right)_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,p\}}}$, $P = \left(d_{ij}\right)_{\substack{i\in\{1,\ldots,n\}\\ j\in\{1,\ldots,n\}}}$ (which can be given or computed from X).
Output: A (adjacency matrix), k (number of groups), Γ (Hubert's statistic).
Algorithm:

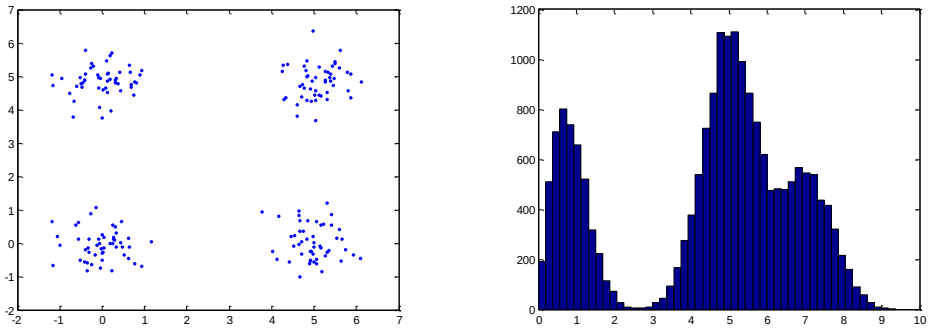
$d_{\max} = \max_{i,j}\left\{d_{ij}\right\}$
for $d \in \left[0, d_{\max}\right]$ (in descending order) do
compute matrix A (if $d_{ij} < d$ then $a_{ij} = 1$ else $a_{ij} = 0$)
compute diagonal matrix D (for every vertex v_i compute $degree(v_i)$)
compute graph Laplacian matrix ($L = D - A$)
compute the number of connected components (number of null eigenvalues of L)
if it's a new partition, compute $\Gamma = -corr\left(P, A\right)$ (validity index)
end for
choose partition for which Γ is bigger (best number of clusters and clustering)

3. Application and Results

The method described in section 2 is applied to two datasets, which are the following: first, a dataset with four groups has been generated (50 objects in each group); second, a well-known dataset is used, which is usually referred to as Iris data.
The method is implemented using MATLAB. The software is available from author at request.

3.1 Simulated data

A two-dimensional set of 200 cases with four different, well separated groups, has been generated to test the method (50 points in each group). The plot of the data and the distribution of dissimilarity values between cases are presented in Figure 1.



(a) (b)

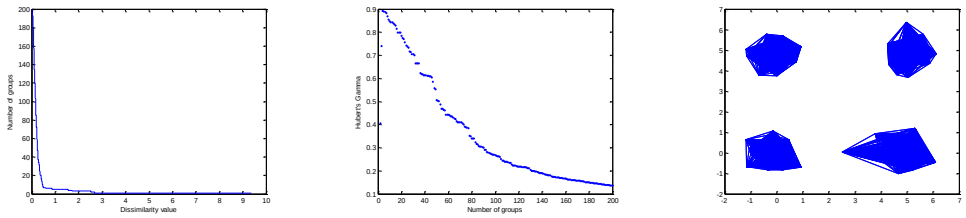
Figure 1: Simulated data. (a) Graphical representation of simulated dataset. (b) Distribution of dissimilarity values in Proximity matrix (Euclidean distance used).

The method has provided an optimal solution where the four real groups have been found. Main results are reported in Table 1.

Number of groups	Distance (d)	Hubert's Gamma (Γ)
2	2.6760	0.4061
3	2.5125	0.7387
4	**1.7343**	**0.8924**
5	1.5545	0.8906
6	0.8590	0.8867

Table 1: Experimental results for simulated data.

If we delete all edges under the value of $d = 1.7343$, the optimal structure reveals that there are four groups in data. No outliers are found, as plotted in Figure 2 (c).
Figure 2 shows the main results of the application of the method to simulated data.



(a) (b) (c)

Figure 2: Simulated data. (a) Number of groups by level of d. (b) Hubert's Γ by number of groups. (c) Final clustering of dataset.

3.2. Iris Data

Iris data is a famous dataset used initially by R.A. Fisher (Fisher, 1936). It has been widely used in taxonomy problems as a test dataset. Four measurements (sepal length and width, and petal length and width) were made on fifty members of each of the three varieties of flower Iris Setosa, Iris Versicolor and Iris Virginica. Iris Setosa is linearly separable from the other two, while there is considerable overlap between Iris Virginica and Iris Versicolor, as can be seen in Figure 3:

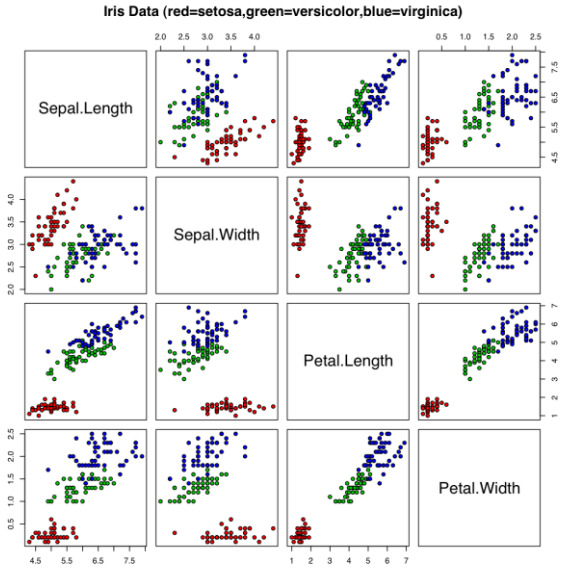


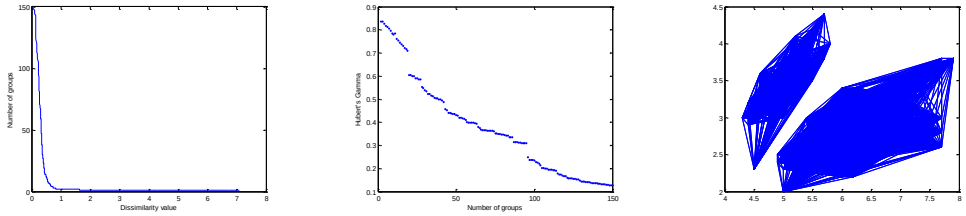
Figure 3: Scatterplot matrix for Iris Data.

The use of Iris dataset in cluster analysis is not very common, as the three species are not linearly separable. Nevertheless, we find some uses of this data set in literature. For example, Davies and Bouldin (1979) propose a clustering of this data set into two groups. They consider the result satisfactory "because in Iris data set two of the three classes have a large overlap". Pal and Biswas (1997) use these data to analyze the validity of clustering. The results obtained with the application of the method proposed in this paper are very similar, as the optimal clustering contains the two usual groups. Main results are reported in Table 2.

Number of groups	Distance (*d*)	Hubert's Gamma (Γ)
2	**1.6401**	**0.8359**
3	0.8185	0.8351
4	0.7348	0.8255
5	0.6481	0.8176

Table 2: Experimental results for Iris data.

If edges over the value of 1.6401 are deleted from graph, the dataset is partitioned into two groups, as shown in Figure 4 (c).



(a) (b) (c)

Figure 4: Iris Data. (a) Number of groups by level of *d*. (b) Hubert's Γ by number of groups. (c) Final clustering of dataset.

4. Discussion and concluding remarks

The aim of this paper was to present a clustering method based on elements of graph theory and to analyze its performance. This method is different to others because it determines simultaneously the number of groups in a dataset and the clustering. All information needed to apply this method is in the dataset, there is no need for additional information. All dissimilarities between objects are taken into account at the beginning, and as a final result it is determined to which cluster every object belongs.
The method has been applied to a self-generated data set, well-separated in four groups. The algorithm has easily discovered automatically the grouping of objects, with no need to specify any extra parameter.
Results obtained for the well-known Iris data set are similar to what can be found in cluster analysis literature. As Iris data is not separable in three groups, the resulting clustering with the proposed method turns out to two clusters.
The data sets analyzed in this paper contain no outliers. Nevertheless, the algorithm proposed can detect outliers in data sets if an object is found to be isolated in a cluster.

References

Davies, D.L.; Bouldin, D.W. (1979) A Cluster Separation Measure, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1 (2), pp. 224-227.

Dubes, R. (1993) *Cluster analysis and related issues*, in Chen, C., Pau, L., Wang, P. (eds.), *Handbook on Pattern Recognition and Computer Vision*, World Scientific Singapore, pp. 3-32.

Fiedler, M. (1973) Algebraic Connectivity of Graphs, *Czechoslovak Mathematical Journal*, 23, pp. 298-305.

Fisher, R.A. (1936) The use of multiple measurements in taxonomic problems, *Annals of eugenics*, 7(2), pp. 179-188.

Gan, G.; Ma, C.; Wu, J. (2007) *Data Clustering: Theory, Algorithms and Applications*, ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA.

Hardy, A. (1996) On the number of clusters. *Computational Statistics & Data Analysis*, 23, pp. 83-96.

Pal, N.R.; Biswas, J. (1997) Cluster Validation using Graph Theoretic Concepts, *Pattern Recognition*, 30 (6), pp. 847-857.

Rokach, L. (2010) *A Survey of Clustering Algorithms*, in Maimon, O.; Rokach, L. (eds.), *Data mining and Knowledge Discovery Handbook*, 2nd Ed., Springer Science+Business Media, LLC, pp. 269-298.

Tan, P.N.; Steinbach, M.; Kumar, V. (2005) *Introduction to Data Mining*. Addison-Wesley.

Von Luxburg, U. (2007) A tutorial on spectral clustering, *Statistics and Computing*, 17 (4), pp. 395-416.