Pattern Recognition Letters 55 (2015) 1–7

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Identifying intervals for hierarchical clustering using the Gershgorin circle theorem ☆

Raghvendra Mall*, Siamak Mehrkanoon, Johan A.K. Suykens

Department of Electrical Engineering, ESAT-STADIUS, Katholieke Universiteit Leuven, Kasteelpark Arenberg 10, B-3001 Leuven, Belgium

ARTICLE INFO

Article history:
Received 10 August 2014
Available online 19 January 2015

Keywords:
Gershgorin circle theorem
k clusters
Eigengap

ABSTRACT

In this paper we present a novel method for unraveling the hierarchical clusters in a given dataset using the Gershgorin circle theorem. The Gershgorin circle theorem provides upper bounds on the eigenvalues of the normalized Laplacian matrix. This can be utilized to determine the ideal range for the number of clusters (k) at different levels of hierarchy in a given dataset. The obtained intervals help to reduce the search space for identifying the ideal value of k at each level. Another advantage is that we don't need to perform the computationally expensive eigen-decomposition step to obtain the eigenvalues and eigenvectors. The intervals provided for k can be considered as input for any spectral clustering method which uses a normalized Laplacian matrix. We show the effectiveness of the method in combination with a spectral clustering method to generate hierarchical clusters for several synthetic and real world datasets.

1. Introduction

Clustering algorithms are widely used tools in fields like data mining, machine learning, graph compression, probability density estimation and many other tasks. The aim of clustering is to organize data into natural groups in a given dataset. Clusters are defined such that the data present within the group are more similar to each other in comparison to the data between clusters. Clusters are ubiquitous and application of clustering algorithms span from domains like market segmentation, biology (taxonomy of plants and animals), libraries (ordering books), WWW (clustering web log data to identify groups) and study of the universe (grouping stars based on similarity) etc. A variety of clustering algorithms exist in literature [1–13] etc. Spectral clustering algorithms [7–9] have become widely popular for clustering data. Spectral clustering methods can handle complex non-linear structure more efficiently than the k-means method. A kernel-based modeling approach to spectral clustering was proposed in Ref. [10] and referred as Kernel spectral clustering (KSC). In this paper we show the effectiveness of the intervals provided by our proposed approach in combination with KSC to obtain inference about the hierarchical structure of a given dataset.

Most clustering algorithms require the end-user to provide the number of clusters (referred as k). This is also applicable for KSC. Though for KSC, we have several model selection methods like balanced line fit (BLF) [10], balanced angular fit (BAF)[11] and Fisher criterion to estimate the number of clusters k which are computationally expensive. However, it is not always obvious to determine the ideal value for k. It is best to choose an ideal value for k based on prior information about the data. But such information is not always available and it makes exploratory data analysis quite difficult particularly when the dimension of the input space is large.

A hierarchical kernel spectral clustering method was proposed in Ref. [14]. In order to determine the optimal number of clusters (k) at a given level of hierarchy the authors in Ref. [14] searched over a grid of values for each kernel parameter σ. They select the value of k corresponding to which the model selection criterion (BLF) is maximum. A disadvantage of this method is that for each level of hierarchy a grid search has to be performed on all the grid values for k. In Ref. [11], the authors showed that the BAF criterion has multiple peaks for different values of k corresponding to a given value of σ. These peaks correspond to optimal value of k at different levels of hierarchy. In this paper we present a novel method to determine the ideal range for k at different levels of hierarchy in a given dataset using the Gershgorin circle theorem [15].

A major advantage of the approach proposed in the paper is that we provide intervals for different levels of hierarchy before applying any clustering algorithm (or using any quality metric) unlike other hierarchical clustering algorithms. The Gershgorin circle theorem provides lower and upper bounds to the eigenvalues of a normalized Laplacian matrix. Using concepts similar to the eigengap, we can use these upper bounds on the eigenvalues to estimate the number of clusters at each

☆ This paper has been recommended for acceptance by M.A. Girolami.

* Corresponding author. Tel.: +32 484287856; fax: +32 1621970.
E-mail address: raghvendra.mall@esat.kuleuven.be (R. Mall).

http://dx.doi.org/10.1016/j.patrec.2014.12.007
0167-8655/

level of hierarchy. Another advantage of this method is that we overcome the computationally expensive eigen-decomposition step. We show the efficiency of the proposed method by providing these discretized intervals (range) as input to KSC for identifying the hierarchy of clusters. These intervals can be used as starting point for any spectral clustering method which works on a normalized Laplacian matrix to identify the k clusters in the given dataset. The method works effectively for several synthetic and real-world datasets as observed from our experiments. Several approaches have been proposed to determine the ideal value of k for a given dataset [7,8,16–25,30]. Most of these methods extend the k-means or expectation maximization and proceed by splitting or merging techniques to increase or decrease the number of clusters respectively.

In this paper we propose a novel method for providing an interval (a range) for the number of clusters (k) in a given dataset. This interval helps to reduce the search space for the ideal value of k. The method uses the Gershgorin circle theorem along with upper bounds on the eigenvalues for this purpose. There are several advantages of the proposed approach. It allows us to identify intervals for the number of clusters (k) at different levels of hierarchy. We overcome the requirement of performing the eigen-decomposition step, thereby reducing the computational cost. There is no underlying assumption or prior knowledge requirement about the data.

2. Proposed method

We consider the normalized Laplacian matrix (L) related to the Random Walk model as defined in Ref. [27]. In this model, the Laplacian matrix is defined as the transition matrix. This can mathematically be represented as $L = D^{-1}S$ where S is the affinity matrix and D is the diagonal degree matrix such that $D_{ii} = \sum_j S_{ij}$. For this model, the highest eigenvalue (equal to 1) has a multiplicity of k in case of k well-separated clusters and a gap between the eigenvalues indicates the existence of clusters. But in real world scenarios there is presence of overlap between the clusters and the eigenvalues deviate from 1. Then it becomes difficult to identify the threshold values to determine the k clusters. Therefore, we utilize the Gershgorin circle theorem to use the upper bounds on the eigenvalues to construct intervals for determining the ranges for the number of clusters (k) at each level of hierarchy in a given dataset. (If we use the normalized Laplacian [$L = I - D^{-1}S$] matrix then it would be required to use the lower bounds on the eigenvalues to construct the intervals). The actual eigenvalues are obtained by performing eigen-decomposition on Laplacian matrix L

$$Lv_j = \lambda_j v_j, \quad j = 1, \ldots, N \tag{1}$$

where N is the number of eigenvalues.

Let $L \in \mathbb{R}^{N\times N}$ be a square matrix which can be decomposed into the sum $L = C + R$ where C is a diagonal matrix and R is a matrix whose diagonal entries are all zero. Let also $c_i = C_{ii}, r_{ij} = R_{ij}$ and $\bar{r}_i = \sum_{j=1}^{N} |r_{ij}|$. Then, according to the Gershgorin circle theorem [15]:

- The ith Gershgorin disc associated to the ith row of L is defined as the interval $I_i = [c_i - \bar{r}_i, c_i + \bar{r}_i]$. The quantities c_i and r_i are respectively referred to as the center and the radius of disc I_i respectively.
- Every eigenvalue of L lies within at least one of the Gershgorin discs I_i.
- The following condition holds:

$$c_j - \bar{r}_j \le \bar{\lambda}_j \le c_j + \bar{r}_j \tag{2}$$

with $\bar{\lambda}_j$ corresponding to disc I_j. For each eigenvalue of L, λ_i, $i = 1, \ldots, N$ there exists an upper bound $\bar{\lambda}_j, j = 1, \ldots, N$ where i need not necessarily be equal to j. Thus, we have $\lambda_i \le \bar{\lambda}_j$.

We are provided with a dataset $\mathcal{D} = \{x_1, x_2, \ldots x_N\}$ where $x_i \in \mathbb{R}^d$. We then construct the affinity matrix S by calculating similarity between each x_i and x_j. Since we use a normalized Laplacian matrix (L) the Gershgorin discs form a set of nested circles and the upper bounds i.e. $\bar{\lambda}_j = c_j + \bar{r}_j$ are all close to 1. However, these $\bar{\lambda}_j$ are more robust and the variations in their values are not as significant as the eigenvalues. It was shown in Ref. [25] that the eigenvalues are positively correlated to the degree distribution in case of real world datasets. This relation can be approximated by a linear function. We empirically observe similar correlations between the degree distribution and these upper bounds i.e. $\bar{\lambda}_j$ generated by the Gershgorin circle theorem. In Ref. [26], the authors perform stability analysis of clustering across multiple levels of hierarchy. They analyze the dynamics of the Potts model and conclude that hierarchical information for multivariate spin configuration could be inferred from spectral significance of a Markov process. In Ref. [26] it was suggested that for every stationary distribution (a level of hierarchy) the spins of the whole system reach the same value. These spin values are dependent on the different eigenvalues and the difference between the eigenvalues of the system. Inspired from this concept we propose a method to use the distance between the upper bounds to determine the intervals to search for optimal values of k for different levels of hierarchy.

We sort these $\bar{\lambda}_j$ in descending order such that $\bar{\lambda}_1 \ge \bar{\lambda}_2 \ge \cdots \ge \bar{\lambda}_N$. Similarly, all the eigenvalues are sorted in descending order such that $\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_N$. The relation $\lambda_1 \le \bar{\lambda}_1$ holds in accordance to the Gershgorin circle theorem. We propose a heuristic i.e. we calculate the distance of each $\bar{\lambda}_j$ from $\bar{\lambda}_1$ to obtain δ_j and maintain this value in a *dist* vector. The distance value is defined as:

$$\delta_j = Dist(\bar{\lambda}_1, \bar{\lambda}_j) \tag{3}$$

where $Dist(\cdot,\cdot)$ is the Euclidean distance function.

We then sort this *dist* vector in descending order. In order to estimate the intervals, we use a concept similar to the notion of eigengap. We first try to locate the number of terms which are exactly the same as $\bar{\lambda}_1$. This can be obtained by calculating the number of terms in the *dist* vector such that $Dist(\bar{\lambda}_1, \bar{\lambda}_j) = 0$. This gives the lower limit for the first interval say $l_1 = n_1$. If there is no $\bar{\lambda}_j$ which is exactly equal to $\bar{\lambda}_1$ then the lower limit for the first interval is 1. We then move to the first term say $\bar{\lambda}_p$ in the sorted *dist* vector which is different from $\bar{\lambda}_1$. We calculate the number of terms say n_2 in the *dist* vector which are at the same distance as $\bar{\lambda}_p$ from $\bar{\lambda}_1$. The upper limit for the first interval is then defined as the sum of the lower limit and the number of terms at the same distance as $\bar{\lambda}_p$ i.e. $u_1 = n_1 + n_2$. This upper limit is also considered as the lower limit for the second interval. We continue this process till we obtain all the intervals. Since we are using the bounds on the eigenvalues ($\bar{\lambda}_j$) instead of the actual eigenvalues (λ_j), it is better to estimate intervals rather than the exact number of clusters. If the length of an interval is say 1 or 2, the search space will be too small. On the other hand, if the length of an interval is too large then we might miss hierarchical structure. So we put a heuristic that the minimum length of an interval should be 3. The intervals provide a hierarchy in a top-down fashion i.e. the number of clusters increases as the level of hierarchy increases. Algorithm 1 provides details of the steps involved to obtain the intervals for each level of hierarchy of a given dataset.

Fig. 1 depicts the steps involved in determining the intervals for estimating the number of clusters (k) at different levels of hierarchy for the R15 [28] dataset. The R15 dataset contains 600 two-dimensional points. There are 15 clusters in this dataset. In Fig. 1(d), we depict the lower limit of the intervals as $l1, l2, l3, l4, l5$ and $l6$ and the upper limit of the intervals as $u1, u2, u3, u4$ and $u5$ respectively. Using these limits the first 5 intervals that we obtain for the R15 dataset are 1–8, 8–12, 12–19, 19–29 and 29–40 respectively. These intervals are obtained using Algorithm 1. From Fig. 1, we show that first we obtain the Gershgorin discs (Fig. 1(a)) which provides us the upper bounds on the eigenvalues. This is followed by the plot of the actual eigenvalues in descending order to show that the actual number of