Expert Systems With Applications 42 (2015) 9489–9498

Contents lists available at ScienceDirect

Expert Systems With Applications

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

A parameter-free similarity graph for spectral clustering

Tülin İnkaya[*]

Uludağ University, Industrial Engineering Department, Görükle Campus, 16059 Bursa, Turkey

ARTICLE INFO

Keywords:
Spectral clustering
Similarity graph
k-nearest neighbor
ε-neighborhood
Fully connected graph

ABSTRACT

Spectral clustering is a popular clustering method due to its simplicity and superior performance in the data sets with non-convex clusters. The method is based on the spectral analysis of a similarity graph. Previous studies show that clustering results are sensitive to the selection of the similarity graph and its parameter(s). In particular, when there are data sets with arbitrary shaped clusters and varying density, it is difficult to determine the proper similarity graph and its parameters without a priori information. To address this issue, we propose a parameter-free similarity graph, namely Density Adaptive Neighborhood (DAN). DAN combines distance, density and connectivity information, and it reflects the local characteristics. We test the performance of DAN with a comprehensive experimental study. We compare k-nearest neighbor (KNN), mutual KNN, ε-neighborhood, fully connected graph, minimum spanning tree, Gabriel graph, and DAN in terms of clustering accuracy. We also examine the robustness of DAN to the number of attributes and the transformations such as decimation and distortion. Our experimental study with various artificial and real data sets shows that DAN improves the spectral clustering results, and it is superior to the competing approaches. Moreover, it facilitates the application of spectral clustering to various domains without a priori information.

1. Introduction

Spectral clustering determines the clusters based on the spectral analysis of a similarity graph. The approach is easy to implement, and it outperforms traditional clustering methods such as k-means algorithm. For this reason, it is one of the widely used clustering algorithms in bioinformatics (Higham, Kalna, & Kibble, 2007), pattern recognition (Vázquez-Martín & Bandera, 2013, Wang, 2008), image segmentation (Zeng, Huang, Kang, & Sang, 2014), and text mining (Dhillon, 2001, He, Qin, & Liu, 2012).

Basically, a spectral clustering algorithm consists of three steps: pre-processing, decomposition, and grouping. In the pre-processing step, a similarity graph and its adjacency matrix are constructed for the data set. In the decomposition step, the representation of the data set is changed using the eigenvectors of the matrix. In the grouping step, clusters are extracted from the new representation. In this study, we focus on the pre-processing step. Our aim is to represent the local characteristics of the data set using a similarity graph. In spectral clustering, we consider three important properties of a similarity graph (Von Luxburg, 2007): (1) The similarity graph should be symmetric and non-negative. (2) The similarity graph should be connected unless the connected components (subclusters) form the target clusters. (3) The similarity graph should be robust.

The most commonly used similarity graphs in the literature are k-nearest neighbor (KNN), mutual KNN, ε-neighborhood, and fully connected graphs (Von Luxburg, 2007). The main idea in these approaches is to represent the local characteristics of the data set using a parameter such as k, ε, and σ. A recent study by Maier, von Luxburg, and Hein (2013) shows that the clustering results depend on the choice of the similarity graph and its parameters. However, proper parameter setting becomes a challenging task for the data sets with arbitrary shaped clusters, varying density, and imbalanced clusters. For instance, KNN may connect the points in different density regions. A similar problem is observed in the ε-neighborhood and fully connected graphs due to the spherical-shaped neighborhoods.

To overcome these limitations a stream of research addresses parameter selection problem for the similarity graph (Nadler & Galun, 2006, Ng, Jordan, & Weiss, 2002, Zelnik-Manor & Perona, 2004, Zhang, Li, & Yu, 2011). Another research stream incorporates the proximity relations to the similarity graph using minimum spanning tree and β–skeleton (Carreira-Perpinan & Zemel, 2005, Correa & Lindstorm, 2012). There are also studies that use k-means, genetic algorithms, and random forests to obtain robust similarity matrices (Beauchemin, 2015, Chrysouli & Tefas, 2015, Zhu, Loy, & Gong, 2014). These approaches provide some improvement, however, they still include parameters to be set properly. Moreover, some of them do not handle the data sets with varying density.

In this study, we propose a parameter-free similarity graph to address the limitations of the aforementioned approaches. We adopt the neighborhood construction (NC) method proposed by

[*] Corresponding author. Tel.: +902242942605; fax: +902242941903.
E-mail address: tinkaya@uludag.edu.tr, tinkaya@gmail.com

http://dx.doi.org/10.1016/j.eswa.2015.07.074
0957-4174/© 2015 Elsevier Ltd. All rights reserved.

İnkaya, Kayalıgil, and Özdemirel (2015) to reflect the local characteristics of the data set. NC yields a unique neighborhood for each point, and the similarity graph generated using NC neighborhoods may be asymmetric. Also, it may include isolated vertices and subgraphs. However, spectral clustering algorithms require symmetric and connected subgraphs. In order to satisfy these properties, we perform additional steps. First, we construct an undirected graph using NC neighborhoods. We call this graph *Density Adaptive Neighborhood* (DAN). Then, we insert edges to DAN if it includes more connected components than the target number of clusters. Finally, we form the weighted adjacency matrix of DAN using Gaussian kernel function. In order to find the clusters, decomposition and grouping steps of any spectral clustering algorithm are applied to the proposed approach. Our comprehensive experimental study with various artificial and real data sets shows the superiority of DAN to competing approaches.

To sum up, our contribution is the development of a preprocessing step for spectral clustering with no a priori information on the data set. The proposed approach includes the construction of a parameter-free similarity graph and its weighted adjacency matrix. It is flexible in the sense that it can be applicable to any spectral clustering algorithm. It works in the data sets with arbitrary shaped clusters and varying density. Moreover, it is robust to the number of attributes and transformations.

The rest of the paper is organized as follows. The related literature is provided in Section 2. We introduce the background information about spectral clustering and similarity graphs in Section 3. The proposed approach is explained in Section 4. The performance of the proposed approach is examined in Section 5. The discussion of the experiments is given in Section 6. Finally, we conclude in Section 7.

2. Literature review

Spectral clustering has its roots in graph partitioning problem. Nascimento and Carvalho (2011), Von Luxburg (2007), and Jia, Ding, Xu, and Nie (2014) provide comprehensive reviews about the spectral clustering algorithms.

The literature about spectral clustering can be classified into two categories (Zhu et al., 2014): (1) The studies that focus on data grouping when a similarity graph is given, and (2) the studies that focus on similarity graph construction when a particular spectral clustering algorithm is used. In the first category, there are several studies that improve the clustering performance. For instance, Liu, Poon, Liu, and Zhang (2014) use latent tree models to find the number of leading eigenvectors and partition the data points. Lu, Fu, and Shu (2014) combine spectral clustering with non-negative matrix factorization, and propose non-negative and sparse spectral clustering algorithm. Xiang and Gong (2008) introduce a novel informative/relevant eigenvector selection algorithm, which determines the number of clusters.

In this study, we address the similarity graph construction problem, so our work is related to the second category. A group of studies in the second category aims to determine the local characteristics of the data set using proper parameter selection. Ng et al. (2002) suggest the execution of spectral clustering algorithm for different values of neighborhood width σ. Then, they pick the one having the least squared intra-cluster distance to the centroid. This method extracts the local characteristics better. However, additional parameters are required, and the computational complexity is high. Zelnik-Manor and Perona (2004) propose the calculation of a local scaling parameter σ_i for each data point instead of a global parameter σ. However, this approach has limitations for the data sets with density variations. Zhang et al. (2011) introduce a local density adaptive similarity measure, namely Common-Near-Neighbor (CNN). CNN uses the local density between two points, and reflects the connectivity by a set of successive points in a dense region. This approach helps scale parameter σ in the Gaussian similarity function. In an alternative scheme, Nadler and Galun (2006) introduce a coherence measure for a set of points in the same cluster. The proposed measure is compared with some threshold values to accept or reject a partition. Although this approach finds the clusters correctly, it is not capable of finding clusters with density variations.

Carreira-Perpinan and Zemel (2005), and Correa and Lindstorm (2012) use proximity graphs to incorporate the connectivity information to the similarity graph. Carreira-Perpinan and Zemel (2005) propose two similarity graphs based on minimum spanning tree (MST). Both graphs are constructed using an ensemble of trees. In the first graph, each point is perturbed using a noise model, and a given number of MSTs are constructed using perturbed versions of the data set. Then, these MSTs are combined to obtain the similarity graph. In the second one, a given number of MSTs are constructed such that the edges in the MSTs are disjoint. Then, the combination of these disjoint MSTs forms the similarity graph. Correa and Lindstorm (2012) introduce an approach that combines β–skeleton (empty region) graph with a local scaling algorithm. The local scaling algorithm uses a diffusion-based mechanism. It starts from an estimate of the local scale, and the local scale is refined for some iterations. Two parameters are used to control the diffusion speed. Although these approaches find arbitrary shaped clusters, density relations among the data points are not reflected to the similarity graphs. Moreover, their performances are sensitive to the proper parameter selection.

A group of studies combine various methods to improve the similarity matrix construction. For example, a recent study by Beauchemin (2015) proposes a density-based similarity matrix construction method based on *k*-means with subbagging. The subbagging procedure increases the density estimate accuracy. However, the proposed approach requires six hyperparameters. Moreover, it has shortcomings when there is manifold proximity in the data set. Zhu et al. (2014) use clustering random forests to obtain a robust similarity matrix. A binary split function is optimized for learning a clustering forest. This also includes two parameters. Chrysouli and Tefas (2015) combine spectral clustering and genetic algorithms (GA). Using GA, they evolve a number of similarity graphs according to the clustering result.

There are also other variants of spectral clustering algorithms. For example, approximate spectral clustering (ASC) is developed for large data sets. ASC works with the representatives of data samples (points), namely prototypes. Hence, the desired similarity matrix should reflect the relations between the data samples and prototypes. Taşdemir (2012) adopts the connectivity graph proposed by Taşdemir and Merényi (2009), and introduces a similarity measure for the vector quantization prototypes, namely CONN. CONN calculates the similarity measure considering the distribution of the data samples in the Voronoi polygons with respect to the prototypes. Taşdemir, Yalçin, and Yildirim (2015) extend this idea and incorporate topology, distance and density information using geodesic-based similarity criteria. Different from these studies, we aim to define the relations among all points in the data set.

In this study, we propose a pre-processing step for spectral clustering, with no a priori information. The proposed approach yields a similarity graph and its weighted adjacency matrix, which can be used with any spectral clustering algorithm. Our work differs from the previous studies in the following sense: (1) It is a parameter-free approach. (2) It reflects the connectivity, density and distance relations among all data points. (3) It works on the data sets not only with convex clusters, but also with clusters having arbitrary shapes and varying density. (4) It is robust to the transformations in the data set.

3. Spectral clustering

In this section, we explain the most commonly used similarity graphs and spectral clustering algorithms in the literature.