Pattern Recognition 48 (2015) 1465–1477

Contents lists available at ScienceDirect

Pattern Recognition

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

Approximate spectral clustering with utilized similarity information using geodesic based hybrid distance measures

Kadim Taşdemir [a,*], Berna Yalçin [b], Isa Yildirim [b]

[a] Department of Computer Engineering, Antalya International University, Universite Caddesi No. 2, Dosemealti, 07190 Antalya, Turkey
[b] Department of Electronics and Communications Engineering, Istanbul Technical University, Ayazaga Kampusu, Ayazaga, Istanbul, Turkey

ARTICLE INFO

Article history:
Received 16 May 2014
Received in revised form
13 October 2014
Accepted 20 October 2014
Available online 24 October 2014

Keywords:
Approximate spectral clustering
Geodesic distances
Hybrid similarity measures
Manifold learning
Cluster validity indices

ABSTRACT

Spectral clustering has been popular thanks to its ability to extract clusters of varying characteristics without using a parametric model in expense of high computational cost required for eigendecomposition of pairwise similarities. In order to utilize its advantages in large datasets where it is infeasible due to its computational burden, approximate spectral clustering (ASC) methods apply spectral clustering on a reduced set of points (data representatives) selected by sampling or quantization. This two-step approach (i.e. finding the representatives and then clustering them) brings new opportunities for precise similarity definition such as manifold based topological relations, data distribution within the Voronoi polyhedra of the representatives, and their geodesic distance information, which are often ignored in similarity definition for ASC. In this study, we propose geodesic based hybrid similarity criteria which enable the use of different types of information for accurate similarity representation in ASC. Despite the fact that geodesic concept has been widely used in clustering, our contribution is the unique way of representing data topology to form geodesic relations and jointly harnessing various information types including topology, distance and density. The proposed criteria are tested using both sampling (selective sampling) and quantization (neural gas and k-means++) approaches. Experiments on artificial datasets, well-known small/medium-size real datasets, and four large datasets (four remote-sensing images), with different types of clusters, show that the proposed geodesic based hybrid similarity criteria outperform traditional similarity criteria in terms of clustering accuracies and several cluster validity indices.

1. Introduction

Spectral methods, recently popular approach in clustering, have a manifold learning algorithm based on eigenvalue decomposition of pairwise similarities of the data points [1–5]. Due to its ability to extract irregularly shaped clusters, its independence from parametric cluster models, and its easy implementation, spectral clustering has been empirically and theoretically supported [6,7] and thus it has been successfully used in various areas such as information retrieval, computer vision, and image processing [8–10]. However, its effective submanifold (cluster) extraction based on eigendecomposition has a drawback of high computational cost ($O(N^3)$, N : number of data points) due to the very same reason. This makes direct use of spectral clustering infeasible for clustering large datasets.

One approach for clustering of large datasets with spectral methods is the use of parallel clustering distributed over many computers [11], to address the memory and computation problem in expense of extra resources. Another approach, which is called approximate spectral clustering (ASC), is to apply spectral clustering on the reduced set of data representatives either selected by a sampling approach or data quantization [12–18]. The ASC methods mainly focus on finding a suitable sampling or quantization method to find the data representatives, with a similarity criterion defined by (Euclidean) distance based Gaussian function. Fowlkes et al. [12] use random sampling based on Nystrom method whereas Wang et al. [15] show that selective sampling is the best sampling method and it has a similar success with k-means quantization. Yan et al. [16] use k-means and random projection trees as quantization methods to conclude experimentally that the best sampling can be achieved by vector quantization with minimum distortion. In addition, there is theoretical justification for using quantization with minimum distortion to determine the data representatives for approximate spectral clustering [19]. Taşdemir [18] compares neural networks (self-organizing maps [20] and neural gas [21]) with k-means and achieves superior ASC accuracies with neural gas quantization. Alternatively, k-means++ [22], a successful variant of k-means with a novel probabilistic approach for initialization, can be a good alternative for quantization in ASC.

[*] Corresponding author. Tel.: +90 553 239 0456; fax : +90 242 245 0010.
E-mail address: kadim.tasdemir@antalya.edu.tr (K. Taşdemir).

http://dx.doi.org/10.1016/j.patcog.2014.10.023
0031-3203/© 2014 Elsevier Ltd. All rights reserved.

Besides making spectral methods feasible for large datasets, the ASC approach enables accurate similarity definitions harnessing different information types on the level of data representatives. For example, the reduced set of representatives efficiently partitions the data space into Voronoi polygons (where each representative is the center), and the data points are distributed to these representatives. This not only provides a data density distribution which may determine separation among submanifolds, but also helps identify topological relations of these representatives with respect to the data manifold. Taşdemir [18] exploits these information to some extent by using CONN similarity defined in [23] (CONN is a weighted adjacency matrix where weights show local data distribution with respect to the data topology) to achieve high clustering accuracies than traditional distance based similarity definition. Moreover, accurate definition of topological relations enables the use of geodesic distances for ASC. Despite being an extensively used approach, geodesic distances are ignored in ASC, mainly due to the difficulty in determining the topological relations required for their truthful calculation with respect to the data manifold. In this study, for ASC, we propose geodesic based hybrid similarities using topological information provided by CONN, traditional (Euclidean) distance and local density information. By utilizing all available information types, the proposed similarity criteria outperform non-geodesic based similarities on a wide selection of datasets.

The paper is outlined as follows. First, we briefly explain approximate spectral clustering in Section 2. Then, we propose our geodesic based similarity definitions in Section 3. We show the outperformance of our geodesic approach in Section 4, using datasets with various characteristics (artificial data with basic clustering challenges, datasets from UCI Machine Learning Repository [24], and real world applications). We conclude and provide open problems in Section 5.

2. Approximate spectral clustering (ASC)

Approximate spectral clustering (ASC) has two steps: (i) selection of data representatives by quantization or sampling; (ii) spectral clustering of the selected data representatives. In the first step, three different approaches (selective sampling, neural gas, and k-means++) will be used. The second step of ASC is the same with the traditional spectral clustering approaches except that the use of representatives in ASC provides new information types (such as data topology and local density) for similarity definition. We briefly explain the ASC algorithm summarized in Fig. 1, the sampling/quantization methods used in this study, and existing similarity definitions for ASC below.

2.1. The ASC algorithm

Being associated with relaxed optimization of graph-cut problems, spectral clustering methods use eigendecomposition of a graph *Laplacian* matrix, L, constructed with respect to some optimization criteria [1–3]. It has been indicated in [4,6] that there is no clear advantage among different spectral methods as long as a normalized graph Laplacian is considered [4,6]. Therefore we utilize the method in [2] for approximate spectral clustering (ASC). For a weighted undirected graph $G=(V,S)$ where the nodes V represent the elements (data points or representatives) to be clustered and the edges S are the pairwise similarities between these elements, Ng et al. [2] define a normalized Laplacian matrix:

$$L_{norm}=D^{-1/2}SD^{-1/2}, \tag{1}$$

based on the similarity matrix S and its diagonal degree matrix D with $d_i=\sum_j s(i,j)$ is the total similarities of the node i. The matrix S can be constructed in various ways [4], whereas it is often based on a Gaussian function of the distances as explained in Section 2.3. By using the spectral clustering algorithm in [2] (steps 2–6 below), an ASC algorithm (Fig. 1) to find k clusters can be summarized as follows:

(1) For a dataset with N samples, find n data representatives either by vector quantization or sampling.
(2) Construct a similarity matrix S showing the pairwise similarities of these n data representatives, based on a user-set similarity criterion.
(3) Calculate the degree matrix D and L_{norm} using the similarity matrix S.
(4) Find the k eigenvectors $\{e_1,e_2,\ldots,e_k\}$ of L_{norm}, associated with the k greatest eigenvalues $\{\lambda_1,\lambda_2,\ldots,\lambda_k\}$.
(5) Construct the $n\times k$ matrix $E=[e_1e_2\ldots e_k]$ and obtain $n\times k$ matrix U by normalizing the rows of E to have norm 1, i.e. $u_{ij}=e_{ij}/\sqrt{\sum_k e_{ik}^2}$.
(6) Cluster the n rows of U with the k-means algorithm into k clusters.
(7) Assign the labels of the n representatives to their corresponding data points.

Note that in this ASC algorithm, steps 2–6 are the spectral clustering algorithm defined in [2] with a difference of using data points directly instead of data representatives in step 2.

2.2. Sampling and quantization methods for ASC

Due to the fact that clustering of large datasets necessitates much computational cost and memory requirement, two-step algorithms, which first reduce the number of data points by producing representatives (prototypes) obtained either by sampling or by quantization, and then cluster the representatives, have been common [21,25–27]. This is particularly important for spectral clustering of large datasets, which requires an eigendecomposition of a similarity matrix (infeasible for such data in terms of computation and memory). Therefore the two-step spectral clustering, namely approximate spectral clustering (ASC) has been used based on various sampling methods and quantization approaches [15–19,12,8]. Selective sampling (SS) is shown to be the optimum sampling method [15–17], whereas quantization is often preferred due to high accuracies and theoretical justification [17,18]. Therefore in this study, we primarily employ neural gas quantization (which was shown outperforming in [18]) and compare it with k-means++ and selective sampling. k-means++ [22] is a variant of k-means which achieves high accuracies based on effective initialization. We briefly explain these three methods below.

2.2.1. Neural gas

The neural gas [28] is a neural learning algorithm which produces topology preserving quantization of the data points. The neural units, which will be the quantization prototypes of the dataset, are randomly initialized. Then for a data point v randomly selected from the dataset M, the best-matching unit (BMU), w_i, is found by the minimum Euclidean distance:

$$\|v-w_i\|\leq\|v-w_j\| \tag{2}$$

The BMU w_i and its neighbor prototypes w_j determined by neighborhood function $h_\tau(w_j)$ are adapted by an iterative learning process:

$$w_j(t+1)=w_j(t)+\alpha(t)h_\tau(w_j)(v-w_j(t)) \tag{3}$$