Pattern Recognition 45 (2012) 3034–3044

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Vector quantization based approximate spectral clustering of large datasets

Kadim Taşdemir*

European Commission Joint Research Centre, Institute for Environment and Sustainability, Via E. Fermi 2749, Ispra (VA), Italy

ARTICLE INFO

Article history:
Received 11 May 2011
Received in revised form
13 January 2012
Accepted 13 February 2012
Available online 24 February 2012

Keywords:
Spectral clustering
Large datasets
Vector quantization
Self-organizing maps
Neural gas
CONN similarity
Connectivity

ABSTRACT

Spectral partitioning, recently popular for unsupervised clustering, is infeasible for large datasets due to its computational complexity and memory requirement. Therefore, approximate spectral clustering of data representatives (selected by various sampling methods) was used. Alternatively, we propose to use neural networks (self-organizing maps and neural gas), which are shown successful in quantization with small distortion, as preliminary sampling for approximate spectral clustering (ASC). We show that they usually outperform *k*-means sampling (which was shown superior to various sampling methods), in terms of clustering accuracy obtained by ASC. More importantly, for quantization based ASC, we introduce a local density-based similarity measure – constructed without any user-set parameter – which achieves accuracies superior to the accuracies of commonly used distance based similarity.

1. Introduction

Unsupervised clustering aims to find distinct groups in a dataset, often without a priori information on their structures. A common approach is to construct parametric models based on known number of clusters. Among them the most popular method is the *k*-means clustering (and its variants) which minimizes the total distances of data samples to their corresponding cluster centroid. Another parametric approach is the use of the expectation-maximization algorithm or Gaussian mixture models, which aims to optimize both the cluster centroids and the cluster variances. However, real datasets often do not fit into parametric models, which in turn requires nonparametric clustering methods [1].

Recently, spectral clustering [2–4], which exploits pairwise similarities of data samples using eigendecomposition of their similarity matrix, has been shown to be successful in several areas such as information retrieval and computer vision [5,6]. It has advantageous properties, such as extraction of irregularly shaped clusters without parametric models and easy implementation, and it has been supported by theoretical and empirical studies [7,8]. Detailed reviews on spectral clustering can be found in [9,10]. For large datasets, however, its use is limited since it is often infeasible due to the computational complexity of $O(N^3)$ and memory requirement of $O(N^2)$ with N being the number of samples to be clustered [11].

In order to apply spectral methods for clustering of large datasets, one approach is to use distributed systems for parallelizing spectral clustering on many computers to overcome the issue of memory use and computational complexity [12]. This in turn requires additional resources that should be scaled according to the size of the dataset. A novel approach, applicable only for segmentation of large images, is to apply spectral clustering to non-overlapping small blocks of the image and combine the resulting partitionings by stochastic ensemble [13]. However, the common naive approach is to reduce the number of data samples using data representatives (either sampled among the data samples or obtained by their quantization), and then apply spectral clustering to those representatives rather than to the data samples directly [11,14–18], producing an *approximate spectral clustering* (ASC). Fowlkes et al. [14] use random selection using Nystrom method, and hence may produce different partitionings at each try. Bezdek et al. [15] use a progressive sampling which has a tendency to over-sample [18], whereas Wang et al. [16,18] use selective sampling. Wang et al. [11] also compare different sampling algorithms for spectral clustering and conclude that selective or *k*-means sampling outperform random sampling approach. Additionally, Yan et al. [17] use *k*-means and random projection trees as sampling methods and show experimentally that vector quantization can be successfully used to select data representatives for fast ASC with slight decrease in clustering accuracy. Moreover, Belabbas and Wolfe [19] provide theoretical justification for using vector quantization to determine the data representatives for approximate spectral clustering.

* Tel.: +39 0 332 78 5040; fax: +39 0 332 78 9029.
E-mail address: kadim.tasdemir@jrc.ec.europa.eu

0031-3203/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2012.02.012

Self-organizing maps (SOMs) [20] and neural gas [21] are two neural networks that can be used for effective vector quantization of large datasets. Contrary to the k-means quantization, which is based on iterative adaptation of the centroids (the best-matching units, BMUs), SOMs and neural gas cooperatively adapt the best-matching units together with their neighbors (determined by a function), to reflect the data topology as faithful as possible with the given number of quantization prototypes. On the one hand, SOMs use a rigid (usually 2D or 3D) grid structure to define the neighborhood relations. This also enables the visualization of high-dimensional data spaces, without dimensionality reduction, since prototypes neighbor in the grid are (ideally) expected to be neighbors in the data space as well. On the other hand, the neural gas defines the neighborhood function in the data space by using the ranking of distances between the prototypes, without any forced layout. Thanks to their quantization based on cooperative adaptation, SOMs and neural gas are successfully used in prototype-based data analysis [22,23]. Our first contribution in this study is to utilize the quantization property of SOMs and neural gas as preprocessors for approximate spectral clustering of large datasets, and show that they are usually superior to k-means quantization, in terms of accuracies achieved by ASC.

In general, another challenge in spectral clustering is to construct the similarity matrix for eigendecomposition. Even though different ways can be used for this matrix [9], a common approach is to define pairwise similarities, $s(i,j)$ s, using a Gaussian function based on the (often Euclidean) distances, $d(x_i,x_j)$, of data samples x_i and x_j, i.e.

$$s(i,j) = e^{-(d(x_i,x_j)^2/2\sigma^2)} \tag{1}$$

where σ is a decaying parameter determining the neighborhood. Alternatively, a recent method [24] defines $s(i,j)$ by including common-near-neighbor, $CNN(i,j)$ (the number of data samples in the intersection of ϵ-neighborhoods of x_i and x_j), as

$$s(i,j) = e^{-(d(x_i,x_j)^2/(2\sigma^2(CNN(i,j)+1)))} \tag{2}$$

and show superior clustering accuracies. However, both approaches requires to set σ which has to be determined properly for the best possible partitioning with spectral clustering. Ref. [3] recommends to use various σ values to find the optimum value whereas [25] uses a cluster ensemble approach to merge partitionings obtained by different σ. Instead, automated setting of σ (different σ_i for each sample x_i) has also been used [26,6,18] by defining σ_i as the distance to the kth nearest neighbor of data sample x_i. However, this approach introduces another parameter to be set by the user, often specific to the dataset [24]. To overcome this challenge for vector quantization based approximate spectral clustering, we define a similarity matrix based on local data distribution without any user-defined parameters, as our second contribution in this study.

The paper is outlined as follows. First, we briefly discuss spectral clustering methods in Section 2; then we describe self-organizing maps and neural gas, which are vector quantization methods for approximate spectral clustering used in this study, in Section 3. In Section 4, we describe our similarity matrix derived from local data distribution. In Section 5, we show the effectiveness of the proposed approaches using three synthetic datasets in [27], six real datasets from UCI Machine Learning Repository [28], and five large datasets. We conclude in Section 6.

2. Spectral clustering

Spectral clustering methods are associated with relaxed optimization of graph-cut problems, using a graph *Laplacian* matrix, L [2–4]. We refer to the tutorials [9,10] (and references therein) for detailed overview of different methods. Below, we describe the method in [3] utilized for this study, since several studies indicate that there is no clear advantage among different spectral methods as long as a normalized graph Laplacian is considered [9,8].

Let $G=(V,S)$ be a weighted, undirected graph with nodes V representing n points in $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}$ to be clustered and edges defined by $n\times n$ similarity matrix S (often constructed by Eq. (1)). Let D be the diagonal matrix denoting the degree of n nodes where $d_i=\sum_j s(i,j)$. Then, clustering of $\mathcal{X}$ can be formulated as a graph-cut problem, which partitions the nodes into two sets $P1$ and $P2=V\backslash P1$, with respect to an optimization function [29]. To achieve balanced cardinality of the resulting partitions $P1$ and $P2$, a popular way is to optimize the normalized cut

$$Ncut(P1,P2) = \frac{\sum_j^n s(i,j)}{\sum_{v_i\in P1} d_i} + \frac{\sum_j^n s(i,j)}{\sum_{v_j\in P2} d_j} \tag{3}$$

Due to high complexity in optimization of graph-cut problems, their optimization is relaxed by spectral graph analysis, introducing the Laplacian matrix, $L=D-S$ (a linear operator on G, based on the similarity matrix S and degree matrix D), and its spectral decomposition [29]. The *Laplacian* matrix, L, is constructed in various ways depending on the approach for graph-cut optimization [9,10]. Ref. [2] shows that the use of eigenvector decomposition (the second smallest eigenvalue and its corresponding eigenvector) of the normalized Laplacian matrix, L_{norm},

$$L_{norm} = D^{-1/2}LD^{-1/2} = D^{-1/2}(D-S)D^{-1/2} = I-D^{-1/2}SD^{-1/2} \tag{4}$$

achieves an approximate solution to the normalized cut (Eq. (3)). To extend the solution for extraction of k clusters, Ng et al. [3] define another normalized Laplacian matrix, $L_{norm[3]}$, based on similarity matrix S

$$L_{norm[3]} = D^{-1/2}SD^{-1/2} \tag{5}$$

and find its k eigenvectors with the k highest eigenvalues. Due to the use of S in $L_{norm[3]}$, the eigenvectors with the k highest eigenvalues are used in clustering, contrary to the use of those with the smallest eigenvalues in [2] which uses L_{norm}. The algorithm of Ng et al. [3] has the following steps:

1. Calculate a similarity matrix S (Eq. (1)), diagonal degree matrix D, and normalized Laplacian $L_{norm[3]}$.
2. Find the k eigenvectors $\{e_1,e_2,\ldots,e_k\}$ of $L_{norm[3]}$, associated with the k highest eigenvalues $\{\lambda_1,\lambda_3,\ldots,\lambda_k\}$.
3. 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. $u_{ij}=e_{ij}/\sqrt{\sum_k e_{ik}^2}$.
4. Cluster the n rows of U with the k-means algorithm into k clusters.

For vector quantization based approximate clustering, we first obtain the quantization prototypes with neural networks (described in the next section), then cluster the quantization prototypes with the above algorithm. Additionally, the similarity matrix is calculated also by local σ_i values [26] and by a density-based similarity measure (CONN) described in Section 4.

3. Vector quantization by neural networks

This section briefly describes the two neural networks, self-organizing maps [20] and neural gas [21], which are used as vector quantization methods for selecting the data representatives to be clustered using spectral methods.