Pattern Recognition Letters 38 (2014) 63–69

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Local information-based fast approximate spectral clustering ☆

Jiangzhong Cao [a,b], Pei Chen [a,*], Qingyun Dai [b], Wing-Kuen Ling [b]

[a] *School of Information Science and Technology, Sun Yat-sen University, Guangzhou 510006, China*
[b] *School of Information Engineering, Guangdong University of Technology, Guangzhou 510006, China*

ARTICLE INFO

Article history:
Received 14 July 2013
Available online 19 November 2013

Keywords:
Spectral clustering
Local information
Sparse affinity graph
Local interpolation

ABSTRACT

Spectral clustering has become one of the most popular clustering approaches in recent years. However, its high computational complexity prevents its application to large-scale datasets. To address this complexity, approximate spectral clustering methods have been proposed. In these methods, computational costs are reduced by using approximation techniques, such as the Nyström method, or by constructing a smaller representative dataset on which spectral clustering is performed. However, the computational efficiency of these approximation methods is achieved at the cost of performance degradation. In this paper, we propose an efficient approximate spectral clustering method in which clustering performance is improved by utilizing local information among the data, while the scalability to the large-scale datasets is retained. Specifically, we improve the approximate spectral clustering method in two aspects. First, a sparse affinity graph is adopted to improve the performance of spectral clustering on the small representative dataset. Second, local interpolation is utilized to improve the extension of the clustering result. Experiments are conducted on several real-world datasets, showing that the proposed method is efficient and outperforms the state-of-the-art approximate spectral clustering algorithms.

1. Introduction

Clustering, or cluster analysis is widely used in many research fields, including pattern recognition, data mining, image processing, and others. Many clustering algorithms have been developed in past decades. Among these, spectral clustering, as a promising method, has recently attracted considerable attention (Alzate and Suykens, 2010; Tasdemir, 2012; Luxburg, 2007; Wang and Dong, 2012). Spectral clustering, which reveals cluster structures of data by using eigenvectors of the Laplacian graph, can stably detect nonconvex patterns and nonlinearly separable clusters (Shi and Malik, 2000; Ng et al., 2002; Guattery and Miller, 1998). It is considered one of the most promising clustering techniques because of its superior performance to traditional clustering algorithms when performed on certain challenging datasets (Verma and Meila, 2003). However, spectral clustering becomes infeasible when applying to large-scale datasets, because its computational complexity increases cubically as the number of data points increases.

Several methods have been developed to apply spectral clustering to large datasets by speeding up the spectral clustering algorithm. These methods can be loosely classified into two types. One type accelerates spectral clustering by reducing the computation of the eigen-decomposition of the Laplacian graph (Fowlkes et al., 2004; Tasdemir, 2012; Wang and Dong, 2012, Shang et al., 2011; Chen and Cai, 2011; Zhang and Kwok, 2009). This method is closely related to low-rank matrix approximations (Williams and Seeger, 2000). The Nyström method, for example, which originated from the numerical solution of continuous eigenfunction problems, has been commonly used in approximate spectral clustering. Fowlkes et al. (2004) first proposed an approximate spectral clustering method based on the Nyström approach. This method interpolates the complete clustering solution by using only a small number of randomly selected samples. Recently, other approximate spectral clustering algorithms based on Nyström methods have been proposed (Tasdemir, 2012; Shang et al., 2011; Zhang and Kwok, 2009). Their differences lie mainly in the sampling strategy, showing that the sampling step is one of the factors that influence the performance of Nyström methods.

The other type of the approximate spectral clustering methods samples a representative data set on which the spectral clustering is performed, and the result is extended to the whole data set (Yan et al. 2009; Shinnou and Sasaki, 2008). Yan et al. (2009) developed a general framework for this type of approximate spectral clustering. They also present two concrete instances under this framework, one based on *K*-means clustering (KASP) and the other based on random projection trees (RASP). KASP is faster than approximate spectral clustering based on the Nyström method, while having comparable accuracy and a significantly smaller memory. In addition, Chen et al. (2011) designed a distributed system for parallelizing spectral clustering where more hardware is

☆ This paper has been recommended for acceptance by M. Kamel.
* Corresponding author. Tel.: +86 20 39332630; fax: +86 20 39943315.
E-mail address: chenpei@mail.sysu.edu.cn (P. Chen).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.005

required. Spectral clustering can be applied to large-scale datasets with these methods however, as mentioned earlier, it is achieved at the cost of performance.

In this paper, we propose an improved approximate spectral clustering method based on local information. In the proposed method, an affinity graph with only local relations is adopted to improve spectral clustering performance on a small representative set, and local interpolation is proposed to improve the extension of the clustering result on the small representative set. The proposed method can obtain good clustering performance and retain scalability to large-scale datasets.

The rest of this paper is organized as follows. The related works on approximate spectral clustering is reviewed in Section 2. In Section 3, we propose an efficient approximate spectral clustering based on local information. Experimental results on several datasets are presented in Section 4. Finally, concluding remarks are provided in Section 5.

2. Approximate spectral clustering

2.1. Spectral clustering

Spectral clustering is a class of methods based on eigen-decompositions of graph affinity matrices. Given a set of data points $S = \{x_1, \ldots, x_n\}$, a weighted graph $G = (V, E)$ is first constructed in which every vertex corresponds to a point in S and each edge is weighted by the similarity between the connected points. The Laplacian graph L (Chung, 1997) is then derived from the adjacency matrix W of G, and the eigenvectors of L are computed. Finally, the traditional K-means method is applied to the low dimensional representations of the original data. There are many spectral clustering algorithms that are based on the above procedures (Luxburg, 2007; Shi and Malik, 2000; Ng et al., 2002).

In this paper, our proposed approximate algorithm is developed in the Jordan-Weiss (NJW) framework (Ng et al., 2002). Therefore, the NJW algorithm is briefly reviewed as Algorithm 1, for the reason of completeness.

Algorithm 1. NJW spectral clustering algorithm

```
Input: Dataset $S = \{x_1, \ldots, x_n\}$ in $\Re^1$ and the number of clusters $k$
Output: $k$-way partition of the input data
(1) Construct the affinity matrix $A$ by the following Gaussian kernel function:
    $A_{ij} = \begin{cases} \exp(\frac{-\|x_i - x_j\|^2}{\delta^2}) & \text{for } i \neq j, \\ 0 & \text{for } i = j, \end{cases}$    (1)
    where $\delta$ is a scale parameter to control how fast the similarity attenuates with the distance between the data points $x_i$ and $x_j$ .
(2) Compute the normalized affinity matrix $\mathbf{L} = \mathbf{D}^{-1/2}\mathbf{A}\ \mathbf{D}^{-1/2}$, where $\boldsymbol{D}$ is the diagonal matrix with $D_{ii} = \sum_{j=1}^{n} A_{ij}$.
(3) Compute the $k$ eigenvectors of $\mathbf{L}$, $v_1, v_2, \ldots, v_k$, which are associated with the $k$ largest eigenvalues, and form the matrix $X = [v_1 v_2, \ldots, v_k]$.
(4) Renormalize each row to form a new matrix $Y \in \Re^{n \times k}$ with $Y_{ij} = X_{ij}/(\sum_j X_{ij}^2)^{1/2}$, so that each row of $\boldsymbol{Y}$ has a unit magnitude.
(5) Treat each row of $\boldsymbol{Y}$ as a point in $\Re^k$ and partition the $n$ points ($n$ rows) into $k$ clusters via a general cluster algorithm, such as the $K$-means algorithm.
(6) Assign the original point $x_i$ to the cluster $c$ if and only if the corresponding row $i$ of the matrix $\boldsymbol{Y}$ is assigned to the cluster $c$.
```

2.2. Fast approximate spectral clustering

In the spectral clustering algorithm above, the major computational burden lies in the construction of the affinity matrix and the computation of the eigenvectors of the Laplace matrix, with a computational complexity of $O(n^2)$ and $O(n^3)$, respectively. Hence, when the number of data points is large, the computational burden of the spectral clustering method becomes unbearable, preventing its application to large-scale datasets.

Several algorithms have been recently proposed to solve this complexity problem by reducing spectral clustering computation. As mentioned earlier, these algorithms can be broadly categorized into two classes. Among these methods, KASP has the advantages of simplicity and speed. In this section, we briefly review the KASP algorithm (Yan et al., 2009), described below as Algorithm 2, because it shares a common framework with the algorithm being proposed. This common framework consists of three steps: (i) constructing a representative set, (ii) performing spectral clustering on the representative set, and (iii) extending the clustering result of the representative set to the whole dataset. These steps influence the performance of approximate spectral clustering in varying degrees. The influence of the first step has already been analyzed by Yan et al. (2009). We aim to improve approximate spectral clustering by enhancing the last two steps, as described in the next section.

Algorithm 2. K-means-based approximate spectral clustering (KASP)

```
Input: Dataset $S = \{x_1, \ldots x_n\}$ in $\Re^1$; the number of clusters, $k$; and the number of representative points, $p$
Output: $k$-way partition of the input data.
(1) Perform the $K$-means clustering with $p$ clusters on the dataset $S$. Use the $p$ centroids of the clusters, denoted by $Y = \{y_1, y_2 \ldots, y_p\}$, as the representative set and construct the correspondence table associating each $x_i$ with the nearest cluster centroid $y_j$.
(2) Perform a spectral clustering algorithm on $Y = \{y_1, y_2 \ldots, y_p\}$ to obtain the $k$-way partition of $Y$.
(3) Assign cluster membership for each $x_i$ by looking up the cluster membership of the corresponding center $y_j$ in the correspondence table.
```

3. Approximate spectral clustering based on local information

In this section, we present an efficient approximate spectral clustering method based on local information. The proposed method is based on two assumptions: (1) The points in the same cluster have more similarity; and (2) The nearby points are likely to have the same label. In other words, the rows of Y in the NJW algorithm that correspond with nearby points are likely to be closer. These assumptions have been applied to many clustering algorithms (Zhou and Bousquet, 2004).

Based on the assumptions above, we improve the approximate spectral algorithm in the following aspects. First, starting from assumption (1), we analyze that the ideal affinity graph for spectral clustering should be sparse; therefore, the approximate ideal affinity graph with local information is adopted in the proposed method to improve performance of spectral clustering on the representative points. Second, based on assumption (2), we propose local interpolation to improve the extension from the clustering result to the whole result. The details are presented in the next subsections.