Pattern Recognition 47 (2014) 418–426

Contents lists available at ScienceDirect

Pattern Recognition

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

Non-negative and sparse spectral clustering

Hongtao Lu [a,b,*], Zhenyong Fu [a,b], Xin Shu [a,b]

[a] Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China
[b] MOE-Microsoft Laboratory for Intelligent Computing and Intelligent Systems, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China

ARTICLE INFO

Article history:
Received 7 December 2012
Received in revised form
24 April 2013
Accepted 4 July 2013
Available online 16 July 2013

Keywords:
Spectral clustering
Non-negative matrix factorization
Ratio cut
Normalized cut
Sparseness

ABSTRACT

Spectral clustering aims to partition a data set into several groups by using the Laplacian of the graph such that data points in the same group are similar while data points in different groups are dissimilar to each other. Spectral clustering is very simple to implement and has many advantages over the traditional clustering algorithms such as *k*-means. Non-negative matrix factorization (NMF) factorizes a non-negative data matrix into a product of two non-negative (lower rank) matrices so as to achieve dimension reduction and part-based data representation. In this work, we proved that the spectral clustering under some conditions is equivalent to NMF. Unlike the previous work, we formulate the spectral clustering as a factorization of data matrix (or scaled data matrix) rather than the symmetrical factorization of the symmetrical pairwise similarity matrix as the previous study did. Under the NMF framework, where regularization can be easily incorporated into the spectral clustering, we propose several non-negative and sparse spectral clustering algorithms. Empirical studies on real world data show much better clustering accuracy of the proposed algorithms than some state-of-the-art methods such as ratio cut and normalized cut spectral clustering and non-negative Laplacian embedding.

1. Introduction

In recent years spectral clustering has become one of the most popular modern clustering methods, e.g., [1,3–6]. Spectral clustering divides a data set into non-overlapped groups such that the data points in same group are similar as much as possible and the data points in different groups are dissimilar as much as possible. Spectral clustering is based on the spectral graph theory [2]. The main tools for spectral clustering are graph Laplacian matrices. There are different graph Laplacian matrices used for spectral clustering in the literature such as unnormalized graph Laplacian [3,4], normalized graph Laplacian [5,6]. The basic steps of the spectral clustering are the first to construct a similarity graph from data and then to compute the first k eigenvectors corresponding to the k smallest eigenvalues of the graph Laplacian matrix which is derived from the similarity graph, and finally use the k-means algorithm to cluster the k dimensional row vectors of the matrix formed by the first k eigenvectors of the Laplacian as columns [1,5,6].

Spectral clustering can be understood from the point of view of the graph cut. For the data given in the form of a similarity graph, one wants to partition the graph into subgraphs such that the edges between different subgraphs have low weights and the edges within a subgraph have high weights. Spectral clustering can be derived as an approximation of graph partition problem which can be solved by minimizing graph cuts. There are different graph cuts used for graph partition such as ratio cut (Rcut) [3], normalized cut (Ncut) [5] and min–max cut (MMcut) [7]. Spectral clustering also has a margin-based perspective [8]. Spectral clustering methods are widely applied in image segmentation [5,9], gene network analysis [10], speech separation [11] and many other fields whenever clustering methods are employed.

Non-negative matrix factorization (NMF) was first introduced into machine learning and pattern recognition communities by Lee and Seung [12,13]. NMF decomposes a non-negative data matrix into a product of two non-negative matrices. The distinguished feature of NMF is the non-negativity of all elements of the matrices involved. This non-negativity is often encountered in real world data such as intensity value of image pixels, document-term matrix, rating matrix, etc. NMF allowing only non-negative factorization makes the data representation to have part-based meaning [12]. Recent extensions of NMF include convex NMF [14], orthogonal NMF [15], convex-hull NMF [16], etc.

NMF has been proved to be closely related to some classical algorithms in machine learning. For example, probabilistic latent semantic indexing (PLSI) has been proved to be equivalent to NMF [17], kernel k-means has also been proved to be equivalent to NMF [18], and it is found that the spectral clustering with Ncut is also equivalent to NMF [19]. A non-negative Laplacian embedding

* Corresponding author at: Department of Computer Science and Engineering, Shanghai Jiao Tong University, 800 Dongchuan Road, Shanghai 200240, China. Tel.: +86 21 34204879.
E-mail address: htlu@sjtu.edu.cn (H. Lu).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.07.003

(equivalent to Rcut spectral clustering) has been proposed and its connection to NMF has been revealed in [20].

In this paper, we only consider spectral clustering methods using Rcut and Ncut as objective functions. We will prove the equivalence between spectral clustering under some conditions and non-negative matrix factorization with proper constraints. Our result is different from that in [19], where the authors proved that spectral clustering with Ncut objective function is equivalent to the symmetrical NMF which factorizes the symmetrical pairwise similarity matrix as a symmetrical product (i.e., product of a non-negative matrix with its transpose). Our result is also different from that of [20] where the authors consider the non-negative Laplacian embedding (which is equivalent to Rcut spectral clustering), which is also to factorize a symmetrical non-negative matrix into a symmetrical product. In the present work, we consider the data matrix itself. Under proper conditions, we prove that a relaxed spectral clustering algorithm with Rcut objective function is equivalent to non-negative factorization of the data matrix into a product of a non-negative matrix with orthogonal columns and another non-negative matrix. For the spectral clustering with Ncut objective function, it is proved to be equivalent to the similar factorization of the normalized data matrix. So the non-negative matrix factorization framework unifies the spectral clustering algorithms using Rcut and Ncut as objective functions.

This non-negative matrix factorization framework provides an insight into connections between spectral clustering and matrix factorization, under this framework, spectral clustering algorithms can benefit from NMF solving techniques. For example, additional constraints can conveniently be incorporated into the framework to derive new spectral clustering algorithms such as non-negative sparse spectral clustering.

2. Spectral clustering

We start with a brief introduction of spectral clustering. Suppose we are given n data points $x_1, x_2, \ldots, x_n$, each is of dimension p, i.e., $x_i \in R^p$, let $X = (x_1, x_2, \ldots, x_n)$ denotes the data matrix of $p \times n$. From the data points, we can construct a similarity graph $G = (V, E)$, where the node set is $V = \{x_1, x_2, \ldots, x_n\}$, and sometimes we also say that the node set is $V = \{1, 2, \ldots, n\}$ without inducing confusion, E is the edge set. In addition, suppose an $n \times n$ matrix W of pairwise similarities (weights) among these n points is also available. The pairwise similarity weights can be computed from the data points. There are some typical methods to construct the similarity graph and compute the similarities between data points, for example, the ϵ−neighborhood graphs, k-nearest neighbor graphs, and fully connected graphs with Gaussian similarity function [1,2,5]. The similarity matrix can also be independent of the node contents and specified in advance as side information [22]. In this paper, without loss of generality, we assume the pairwise similarity being nonnegative, i.e., $w_{ij} \geq 0$, and the similarity graph is undirected and symmetric, this means $w_{ij} = w_{ji}$. Specifically, we consider the similarity matrix based on the inner product, i.e., $W = X^T X$.

2.1. Graph Laplacian

For a similarity graph $G = (V, E)$ with weight matrix W, the degree d_i of node i is

$$d_i = \sum_j w_{ij} \tag{1}$$

The degree matrix D is defined as the diagonal matrix with the degrees $d_1, d_2, \ldots, d_n$ on its diagonal, i.e., $D = \mathrm{diag}(d_1, d_2, \ldots, d_n)$. The unnormalized graph Laplacian matrix is defined as

$$L = D - W. \tag{2}$$

It can be proved that L is symmetric and semi-positive definite, and for any vector $x \in R^n$ [1]

$$x^T L x = \frac{1}{2} \sum_{i,j=1}^{n} (x_i - x_j)^2 w_{ij}. \tag{3}$$

The symmetric normalized graph Laplacian matrix L_{sym} is defined as

$$L_{\mathrm{sym}} = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2} \tag{4}$$

where I is the identity matrix. Similarly

$$x^T L_{\mathrm{sym}} x = \frac{1}{2} \sum_{i,j=1}^{n} \left(\frac{x_i}{\sqrt{d_i}} - \frac{x_j}{\sqrt{d_j}} \right)^2 w_{ij}. \tag{5}$$

2.2. Graph cut

The spectral clustering of data points can be interpreted by the partitioning of the similarity graph into several non-overlapping parts and the problem can be solved by the graph cut approaches. Graph cut algorithms divide a graph into groups such that edges between different groups have low weights and edges within each group have high weights. This can be formulated as a minimization problem of an appropriate graph cut objective function. The first attempt is to directly minimize the cut $s(A, B)$, between two partitions A and B, where the cut $s(A, B)$ is defined as

$$s(A, B) = \sum_{i \in A} \sum_{j \in B} w_{ij}. \tag{6}$$

However, this mincut algorithm is likely to divide out small subgraphs, making the partition severely unbalanced. The ratio cut (Rcut) J_{rc} is one of such objective functions proposed to handle this problem [21]

$$J_{\mathrm{rc}} = \frac{s(A, B)}{|A|} + \frac{s(A, B)}{|B|} \tag{7}$$

where $|A|$ means the cardinality of A. The graph cut problem based on Rcut is solved by minimizing J_{rc} with respect to A and B. Shi and Malik [5] proposed another objective function: the normalized cut (Ncut) J_{nc}:

$$J_{\mathrm{nc}} = \frac{s(A, B)}{d(A)} + \frac{s(A, B)}{d(B)} \tag{8}$$

where $d(A) = \sum_{i \in A} d_i$ is the sum of node degrees of A.

Chan et al. [4] generalized the two-way ratio cut of (7) to multi-way ratio cut where the nodes of graph G are divided into K disjoint groups C_p by minimizing the following multi-way ratio cut objective function:

$$J_{\mathrm{rc}} = \sum_{1 \leq p < q \leq K} \left\{ \frac{s(C_p, C_q)}{|C_p|} + \frac{s(C_p, C_q)}{|C_q|} \right\} \tag{9}$$

Let h_l be the indicator vector for cluster C_l, i.e., $h_l(i) = 1$ if $x_i \in C_l$, otherwise $h_l(i) = 0$, then $|C_l| = h_l^T h_l$. It is easy to verify that

$$J_{\mathrm{rc}} = \sum_{l=1}^{K} \frac{s(C_l, \overline{C}_l)}{|C_l|} \tag{10}$$

where $s(C_l, \overline{C}_l)$ is the cut between C_l and its complement $\overline{C}_l = V - C_l$. $s(C_l, \overline{C}_l)$ can be expressed as

$$s(C_l, \overline{C}_l) = h_l^T (D - W) h_l \tag{11}$$

If we define the $n \times K$ cluster indicator matrix H as

$$H = \left(\frac{h_1}{\|h_1\|}, \frac{h_2}{\|h_2\|}, \ldots, \frac{h_K}{\|h_K\|} \right) \tag{12}$$