Pattern Recognition 52 (2016) 274–286

Contents lists available at ScienceDirect

Pattern Recognition

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

Iterative ensemble normalized cuts

Li He *, Hong Zhang

Department of Computing Science, University of Alberta, 2-21 Athabasca Hall, Edmonton, Alberta, Canada T6G 2E8

ARTICLE INFO

Article history:
Received 2 December 2014
Received in revised form
7 October 2015
Accepted 28 October 2015
Available online 10 November 2015

Keywords:
Iterative ensemble NCut
Gap-normalized distance
Spectral clustering
Image segmentation

ABSTRACT

A fast spectral clustering method is proposed. Eigenvectors used in NCut are studied as the gap-normalized distances defined in this paper. The out-of-sample extensions of NCut are derived by extending the gap-normalized distances to new data, which is interestingly found to be perfectly matched with the Nyström-based eigenfunction approximation. From the view of gap-normalized distance, the ensemble NCut method is built by assembling distances of small groups to learn the partitions of the entire dataset. By iteratively calling such assembly, the iterative ensemble NCut method is proposed. Experiments on real world datasets and the image segmentation tasks show that, compared with the state-of-the-art, the proposed IENCut method produces improved clustering quality although this improvement may sometimes come at the expense of increased processing time.

1. Introduction

In the last decade, spectral clustering has been successfully studied in the machine learning and data mining communities. As the starting point of many spectral clustering algorithms, the Normalized Cuts (NCut) algorithm of Shi and Malik [1] is widely used in image segmentation, face recognition [2], document clustering [3], and unusual activity detection [4]. Whereas many clustering methods are strongly attached to the convex region in Euclidean space, NCut is more flexible in terms of handling data with a wider range of geometries. It often shows superior empirical performance when compared with competing algorithms such as k-means or EM, which sometimes fall into the local optimal solutions on non-convex data.

Despite its advantages, NCut is not widely viewed as a competitor for large-scale clustering problems, an area that is usually dominated by classical algorithms such as hierarchical clustering and k-means. This is due to the fact that when handling a dataset with N observations, NCut needs to compute the eigenvectors of the $N \times N$ affinity matrix, an operation with a complexity of $O(N^3)$ and memory space of $O(N^2)$ in general. For a moderately large N, both the time cost and memory requirement prevent NCut to be practical.

Our focus here is to extend NCut to deal with large-scale applications in an efficient way. As in many other situations in spectral clustering, the eigendecomposition on the entire data is the computational bottleneck for NCut. We aim to find an effective scheme that reduces the data size for decomposition, and an approximation to the optimal solution is also desired. In this paper, we propose an ensemble NCut (ENCut) method and its iterative version (IENCut) for large-scale data. The proposed method is based on a geometric perspective of NCut and employs a linear combination of embeddings from different training sets to pursue the solution of NCut problem. Our main contributions include the following:

(1) We extend the gap perspective of NCut in [5], and reveal that elements in eigenvectors used in NCut are naturally the weighted distances of data to a series of orthogonal planes, after gap normalization. Based on this view of NCut, the gap-normalized distance is defined. Furthermore, the out-of-sample extensions from the gap-normalized distance and from the Nyström eigenfunction approximation are found perfectly matched.

(2) Based on the gap-normalized distance, we propose a fast version of NCut, the ensemble NCut, which runs NCut on several small groups of data and assembles the gap-normalized distances. An iterative implementation of ENCut, called IENCut, is proposed to efficiently solve ENCut.

(3) In order to meet the cluster-balancing constraint in IENCut, we use a projection matrix to force the eigenvectors to be orthogonal to the mean of data, as what NCut requires. To handle the problem of uncertain directions of cutting planes learned from groups, we use the sign alignment to adjust the signs of eigenvectors of each group.

The rest of this paper is organized as follows. Section 2 briefly outlines the basic Normalized Cuts algorithm. Section 3 discusses related work in NCut for large-scale tasks. Section 4 describes the gap-normalized distance from the gap perspective and its out-of-sample solutions. Section 5 introduces the ensemble NCut and its iterative version. Experimental results are presented in Section 6, and conclusions are shown in Section 7.

* Corresponding author. Tel.: +1 780 492 7188; fax: +1 780 492 1071.
E-mail addresses: lhe2@ualberta.ca (L. He), zhang@cs.ualberta.ca (H. Zhang).

http://dx.doi.org/10.1016/j.patcog.2015.10.019
0031-3203/

2. Normalized cuts

Like other clustering algorithms, NCut tries to partition data into clusters, so that members within the same cluster are similar to each other and members in different clusters are dissimilar. Such partition is based on the eigendecomposition of the $N \times N$ normalized affinity matrix with elements of the pairwise affinities, where N denotes the number of data. The eigenvectors of the matrix induce the embeddings of the data in a low-dimensional space, where data are easy to partition by a simple clustering method (such as the k-means). More details about NCut will be discussed later.

Given a set of data points $\mathbf{x} = \{\mathbf{x}(i) | \mathbf{x}(i) \in \Re^d, i = 1, 2, \ldots, N\}$, the affinity matrix $\mathbf{K}$ is defined by $K_{ij} = k(\mathbf{x}(i), \mathbf{x}(j))$, where $k(\mathbf{x}, \mathbf{y})$ is a distance (or similarity) measure which is positive definite and, in many cases, symmetric. A common choice for $k(\mathbf{x}, \mathbf{y})$ is the Gaussian kernel, $k(\mathbf{x}(i), \mathbf{x}(j)) = \exp(-\|\mathbf{x}(i) - \mathbf{x}(j)\|^2/\sigma^2)$, where $\|\cdot\|$ denotes the L^2 norm. In graph theory, the affinity matrix $\mathbf{K}$ defines the weights on the edges of a fully connected graph in which each node corresponds to a data $\mathbf{x}(i)$ and K_{ij} the edge between node $\mathbf{x}(i)$ and $\mathbf{x}(j)$.

A binary partition of data, for example, cuts the graph into two disjoint sets, A and B, with binary label $y_i \in \{-1, +1\}$. NCut refers to the cost function $cut(A, B) = \sum_{y_i > 0} \sum_{y_j < 0} -y_i y_j K_{ij}$, defined by the sum of the pairwise weights that link nodes in different sets. Denote the degree matrix $\mathbf{D} = diag\{d_1, d_2, \ldots, d_N\}$ where $d_i = \sum_j K_{ij}$, the sum of the ith row of $\mathbf{K}$. Let $vol(A) = \sum_{i \in A} d_i$ be the sum weight of all data in set A. The goal of NCut is to find the optimal labels $\{y_i\}_{i=1}^N \in \{-1, +1\}$ that minimize the normalized cost function:

$$NCut(A, B) = \frac{cut(A, B)}{vol(A)} + \frac{cut(B, A)}{vol(B)} \tag{1}$$

However, minimizing (1) by binary labels is NP-complete [1]. To handle this problem, [1] proposes to embed labels y_i in real value domain for an approximate solution. The embedded labels, $y_i \in [-1, +1]$, are also known as the embeddings of data.

For multi-way partition, one can refer to the recursive two-way NCut or the K-way NCut with multiple eigenvectors [1]. In recursive two-way NCut, the bipartition of data by (1) is recursively called until the stop criterion, such as the desired NCut cost value, is met. For K-way NCut, the leading l eigenvectors are used as the embeddings of all data, i.e., each data is embedded into a l-dimensional space for clustering. We adapt the K-way NCut in our paper for multi-way NCut partition.

After some manipulations, the problem can be rewritten as

$$\begin{aligned} &\underset{\mathbf{y}}{\operatorname{argmin}} \frac{\mathbf{y}^T(\mathbf{D} - \mathbf{K})\mathbf{y}}{\mathbf{y}^T\mathbf{D}\mathbf{y}} \\ &\text{s.t. } \mathbf{y}^T\mathbf{D}\mathbf{1} = 0 \end{aligned} \tag{2}$$

where $\mathbf{1}$ is the column vector with all elements equal to unity, l is the number of clusters and $\mathbf{y}$ is the real-valued label vector. To solve (2), it is equivalent to compute the largest l eigenvectors $\mathbf{v}$ of the system [5]

$$\mathbf{D}^{-1/2}\mathbf{K}\mathbf{D}^{-1/2}\mathbf{v} = \lambda\mathbf{v} \tag{3}$$

The embeddings (labels) $\mathbf{E}_{NCut} = \mathbf{y}$ are given by

$$\mathbf{E}_{NCut} = \mathbf{D}^{-1/2}\mathbf{v} \tag{4}$$

3. Related works on large-scale clustering

There are several options that can be considered for the extension of NCut to large-scale data. One option is to replace the original data with a small number of points, i.e., the representatives, which intend to reveal the data structure as a small-scale clustering task. The KASP method in [6] runs the k-means on the raw data and picks up the representatives as the centroids of each k-means-based clusters. Then it performs spectral clustering on the obtained representatives and distributes the label of each data point corresponding to its nearest neighbor in cluster centroids. Sakai and Imiya [7] proposed another way which is based on the random projection and sampling method. Chen and Cai [8] represented the data as the linear combination of representatives (the landmarks), and replaced the original affinity matrix by the representation matrix for eigendecomposition. More recently, Shang et al. proposed a fast affinity propagation (FAP) method. This method first makes use of the sparse affinity propagation [9] to select the representative candidates then refines them. Finally, FAP assigns the labels of data corresponding to their representatives. Yu et al. [10] adopted the self-organizing map (SOM) approach to select the set of representatives. Then a couple of matrices, the adjacency matrix $\mathbf{A}$ and the attraction matrix $\mathbf{M}$, are used in Yu's work, where $\mathbf{A}$ is related to the Euclidean distance matrix of representatives and $\mathbf{M}$ stands for the similarities of raw data to the representatives. Instead of the affinity matrix used in NCut, Yu ran NCut on a new matrix, which is a linear weighted combination of $\mathbf{A}$ and $\mathbf{M}$.

Since finding the optimal binary-labeled NCut has proven to be NP-hard, several researchers focus on the spectral relaxation problem. Bie and Cristianini [11] devised an efficient NCut relaxation to a semi-definite program (SDP). Instead of binary labels, say $\mathbf{y} = \{-1, 1\}$, of data, SDP employs the label matrix Γ, which is related to the NCut relaxation matrix $\mathbf{y}\mathbf{y}^T$. With the constraints of semi-definite and unit diagonal, SDP attempts to solve the optimal Γ and subsequently generates the label vector $\mathbf{y}$ from Γ. The complexity of SDP is reduced to roughly $O(N^{2.5})$. SDP uses the 'subspace trick' to take into account the consistency of similar data. Coleman et al. [12] improves the subspace trick by correcting the inconsistent data as an optimal solution to the 2-correlation clustering problem.

Power iteration method is also a popular choice for accelerating NCut. Starting from a randomly selected vector, this kind of methods iteratively updates the approximated eigenvector until a desired approximation accuracy is met. Xu et al. [13] proposed a fast NCut method with linear constraints. In [13] the power iteration algorithm is used for updating $\mathbf{u}$, the constraint component of eigenvectors. Xu's work also shows that the solution from their method is the global solution to the constraint eigendecomposition problem. Instead of using eigenvectors individually, Lin and Cohen [14] proposed to use a linear combination of eigenvectors when the eigengap of the affinity matrix is large enough. In Lin's work, a single vector, the weighted sum of eigenvectors, is used and this vector is shown to be block-wise.

The Nyström method may be the most common approach pursued in the literature. This approach exploits the low-rank approximations to the affinity matrix, and builds the eigenvectors from eigenfunctions learned on a small set of data. Despite several related applications of the Nyström method in machine learning, mostly on the sped-up eigendecomposition of the Gram matrix tasks [15,16], its application to spectral clustering begins with the out-of-sample work in [17]. Works in [17] introduce the Nyström method to learn the underlining eigenfunction of a kernel and approximates the eigenvectors of entire data from a small set of samples. Zhang and Kwok [18] proposed the weighted Nyström spectral clustering method (WNSP) for clustering. In WNSP, the volume-weighted kernel matrix is used for eigendecomposition and the Nyström is employed for extending the eigenvectors from landmarks to other data. WNSP uses k-means on input data first to find the landmarks and takes the size of each cluster as the weight. It is also found in Zhang's work that using the k-means centers as the landmarks could depress the upper bound of the approximation error between the accurate kernel matrix and the landmark-based approximated kernel matrix.