Pattern Recognition Letters 68 (2015) 90–96

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

On affinity matrix normalization for graph cuts and spectral clustering ☆

M. Beauchemin*

Natural Resources Canada, Canada Centre for Remote Sensing, 560 Rochester Street, Ottawa K1A 0E4, Canada

ARTICLE INFO

Article history:
Received 15 July 2014
Available online 3 September 2015

Keywords:
Affinity matrix
Spectral clustering
Kernel *k*-means
Feature space

ABSTRACT

Graph-based spectral clustering algorithms involve the analysis of an affinity matrix. The latter defines the pairwise similarity relations among data points. Popular graph partitioning algorithms typically involve a normalization step that reflects itself onto an affinity matrix normalization step in spectral clustering algorithms. In this paper, we show that affinity matrix normalization with constant row/column sum guarantees the invariance of the size-weighted sum of the between- and within-cluster graph association; a property conceptually equivalent to the data variance decomposition exploited by the standard *k*-means algorithm. From this observation, we demonstrate that the solution of numerous spectral clustering methods can be obtained using the standard graph ratio cut objective function. We have identified in the literature 7 such affinity matrix normalization schemes relevant to spectral clustering. Clustering experiments performed with these 7 normalization schemes on 17 benchmark datasets are presented. As a general rule, it is observed that the appropriate normalization method depends on the dataset. A geometric interpretation in the feature space (FS) of such a normalization scheme for *k*-way spectral clustering is also presented.

1. Introduction

Maximizing the within-cluster similarities while simultaneously minimizing the between-cluster similarities is an approach playing a key role in several popular unsupervised clustering algorithms. For example, minimizing the within-cluster scatter at the base of the well-known *k*-means algorithm is equivalent to maximizing the between-cluster scatter because the within/between scatter sum is an invariant, irrespective of how data are partitioned [10]. The same property, hereafter referred to as minimum/maximum duality optimization, applies for kernel *k*-means [13]. Yet, an analogous minimum/maximum duality exists for the minimum cut and the normalized cut (Ncut) graph partitioning algorithms: both techniques minimize the between-cluster similarities while simultaneously maximizing the within-cluster similarities [22]. Recently, Dhillon et al. [5] introduced a general weighted graph cuts/association formalism incorporating existing graph partitioning algorithms that are based on different cut-normalizing terms (e.g. ratio cut, ratio association, normalized cut). Their study reveals the equivalence between the objective function of spectral clustering and kernel *k*-means. Spectral clustering algorithms for graph partitioning have shown good performance to solve graph partitioning problems [12]. In particular, Dhillon et al. [5] showed that maximizing a general weighted graph association is equivalent to minimizing a general weighted graph cut on some specific matrices comprising the affinity matrix, its graph Laplacian and the matrix of general weights (more details in Section 3). The latter reflects again a minimum/maximum duality optimization criterion for clustering.

From a different perspective, Zass and Shashua [23] demonstrated that the solution of kernel *k*-means objective function is optimized when the input affinity matrix[1] is doubly stochastic, that is each row and column of the input affinity matrix is normalized to sum to one. An affinity matrix provides a measure of the degree of similarity between every pair of points in a dataset. Affinity matrix normalization is defined in this paper as operations applied on the affinity matrix and transforming it into a matrix with constant sum for rows or columns or both of them. Zass and Shashua [23] noticed that the ratio cut and normalized cut algorithms, although not using doubly stochastic matrices, employ the closest doubly stochastic matrix to a given input affinity matrix according to some error measures (ℓ_1 for ratio cut and relative entropy for normalized cut). The closest doubly stochastic matrix must be non-negative and symmetric. Both ratio and normalized cut involve thus an implicit normalization step (see also Table 1, [3]). From these observations, Zass and Shashua argued that a doubly stochastic approximation of the input affinity matrix should be pursued for spectral clustering. Following that reasoning, they proposed the construction of a doubly stochastic (normalized) affinity matrix based on the Frobenius error norm. In a similar way, Wang et al. [20] showed that using the Kullback–Leibler divergence

☆ This paper has been recommended for acceptance by Andrea Torsello.

* Tel.: +1 613 759 6449; fax: +1 613 759 6344.
E-mail address: mario.beauchemin@nrcan.gc.ca

[1] In this paper, we use the terms affinity and similarity matrix interchangeably.

http://dx.doi.org/10.1016/j.patrec.2015.08.020
0167-8655/

Table 1
Normalization schemes resulting in constant row/column sum when applied on a given affinity matrix $\mathbf{U}$.

Normalization operations applied on $\mathbf{U}$	Row/col. sums $\overline{\mathbf{U}}\mathbf{1}$, $\overline{\mathbf{U}}^{\mathrm{T}}\mathbf{1}$	References
$\overline{\mathbf{U}}_{\mathrm{L}} = \mathbf{L} = \mathbf{D} - \mathbf{U}$	$\mathbf{0}$	Hagen and Kahng [9]
$\overline{\mathbf{U}}_{\mathrm{Zj}} = [\mathbf{I} - \mathbf{D}^{-1}\mathbf{U}]^{\mathrm{T}}\,[\,-\mathbf{D}^{-1}\mathbf{U}]$	$\mathbf{0}$	Zhang and Jordan [24]
$\overline{\mathbf{U}}_{\mathrm{RR}} = \mathbf{U} - \frac{\mathbf{U}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{U}}{\mathbf{1}^{\mathrm{T}}\mathbf{U}\mathbf{1}}$	$\mathbf{0}$	Rahimi and Recht [16]
$\overline{\mathbf{U}}_{\mathrm{FP\_mod}} = [\mathbf{I} - t^{-1}\mathbf{U}']^{T}[\mathbf{I} - t^{-1}\mathbf{U}']$, where $[\mathbf{U}']_{ij} = \exp[-(\mathbf{x}_i - \mathbf{x}_j)^2/(2{\sigma_i}^2)]$ and $\forall_i$: $\sum_{j\in C} U'_{it} = \tau$, $\tau -> -0$; σ_i chosen such that each row sum $= \tau$	$\mathbf{0}$	Modified from Fischer and Poland [6]
$\overline{\mathbf{U}}_{\mathrm{Ls\_iter}}$, iterates $\mathbf{U}^{(t+1)} = [\mathbf{D}^{-1/2}\mathbf{U}^{(t)}\mathbf{D}^{-1/2}]$	$\lim_{t\to\infty} \mathbf{U}^{(t)} = \mathbf{1}$	Zass and Shashua [23], Ng et al. [15]
$\overline{\mathbf{U}}_{\mathrm{ZSW}}$, iterates $\mathbf{U}^{(t+1)} = [\mathbf{U}^{(t)} + N^{-1}(\mathbf{I} - \mathbf{U}^{(t)} + N^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{U}^{(t)})\mathbf{1}\mathbf{1}^{\mathrm{T}} - N^{-1}\mathbf{1}\mathbf{1}^{\mathrm{T}}\mathbf{U}^{(t)}]^{+}$	$\lim_{t\to\infty} \mathbf{U}^{(t)} = \mathbf{1}$	Zass and Shashua [23], Wang et al. [20]
$\overline{\mathbf{U}}_{\mathrm{SK}}$, iterates $\mathbf{U}^{(t+1)} = [(U_{ij}/\sum_i U_{ij})^{(t)} \leftarrow (U_{ij}/\sum_j U_{ij})^{(t)}]$	$\lim_{t\to\infty} \mathbf{U}^{(t)} = \mathbf{1}\mathbf{1}$	Sinkhorn and Knopp [18]

Table 2
Summary of the datasets used for the experiments. k = number of groups, N = number of samples, and d = number of features with the reduced dimension in brackets.

Dataset	k	N	d (d_{PCA})
Iris	3	150	4
Wine	3	178	13
Crabs	2	200	5
Sonar	2	208	60
Glass	6	214	9
House	2	232	16
Spectf_Heart	2	267	44
Ionoshere	2	351	34
Std-yeast	5	384	17
Balance	3	625	4
Pima	2	768	8
OptDigits	10	1000	64
Cancer	2	683	9
Dermatology	6	358	34
Coil20	20	1440	1024 (6)
Leukemia	2	72	7129 (6)
ZipCode5-9	5	3051	256 (6)

to approximate the affinity matrix lead to the Sinkhorn and Knopp algorithm for doubly stochastic affinity matrix normalization [18]. Yet, implicit normalization schemes for spectral clustering also exist in algorithms such as the average gap algorithm [16] or the use of an alternative choice for the unnomalized graph Laplacian in ratio cut [24]. The latter two approaches result in a normalization of the input affinity matrix where the row/column sums are equal to zero. The third column of Table 1 provides a list of works involving affinity matrix normalization with constant row/column sums.

This brief survey shows that the normalization of affinity matrices leading to constant row/column sums plays a central role in the solution of numerous graph clustering algorithms. Notice that the lack of normalization generally leads to inferior clustering performances (see, e.g., Tables 2 and 3 of [21]. The fact that there exist connections between the weighted versions of graph clustering and kernel k-means objective functions suggested us to examine the role of affinity matrix normalization from the minimum/maximum duality optimization property perspective.

In this paper, we revisit the minimum/maximum duality optimization property in graph clustering, focusing on the constraints required for affinity matrices to preserve the invariance of within- and between-cluster similarities sum.

The main contributions are, in order of presentation: (i) we establish a relationship with invariant property regarding cluster's data assignment for graph partition algorithms, conceptually equivalent to the data variance decomposition exploited by the standard k-means algorithm; the relationship holds under the condition that the affinity matrix is normalized in such a way that the sum of each row and each column is equal to a same constant, (ii) we show that the solution of numerous spectral clustering methods can be obtained using the standard graph ratio cut objective function for such normalized affinity matrices, (iii) we propose a modified version of the Fisher and Poland [6] affinity matrix construction method that significantly improves clustering performance over the original formulation, and (iv) we provide a geometric interpretation in the feature space of such a normalization scheme for k-way spectral clustering.

The paper is organized as follows. Relevant material from graph partitioning, spectral clustering and kernel k-means is reviewed in Section 2. The main relationships and their significances are derived in Section 3. The feature space interpretation is detailed in Section 4. Section 5 presents and discusses experimental results illustrating the findings of the previous sections. The conclusion is given in Section 6.

2. Preliminaries

In this section, we introduce the notation and basics of relevant algorithms considered in this work: graph partitioning, affinity matrix, spectral clustering, affinity matrix normalization and kernel k-means.

2.1. Graph partitioning

Consider a dataset $\mathbf{X} = \{\mathbf{x}_1,\ldots, \mathbf{x}_N\}$ where $\mathbf{x}_i$ is a vector in a d-dimensional space. Let $G = (V, E, \mathbf{U})$ be an undirected weighted graph where $V = \{V_1,\ldots, V_N\}$ is the set of all nodes (or vertices) corresponding to individual data $\mathbf{x}_i$ and E is the set of edges connecting all pairs of nodes (V_i, V_j). The affinity matrix $\mathbf{U} = [U_{ij}]_{N\times N}$, $U_{ij} \geq 0$, is a measure of the edge weights: it is a matrix quantifying the similarity between points $\mathbf{x}_i$, $\mathbf{x}_j$. Many ways exist for constructing pairwise similarities from a set of data points. The most common ones include ε-neighborhood graph, k-nearest neighbor graph and similarity functions [19]. Co-association matrices constructed from the combination of multiple partitioning algorithms, e.g. Fred and Jain [7], also possess affinity matrix properties.

The clustering problem consists in partitioning V into K disjoints subsets V_i, $i = 1,\ldots, K$, with $V = \bigcup_{i=1}^{K} V_i$ and $V_i \cap V_j = \emptyset$ for $i \neq j$, so that the similarity among the nodes in each individual set is high and the similarity between the different subsets is low. Consider first the bipartite case, $k = 2$, and denote two disjoint sets A and B with $A \cup B = V$ and $A \cap B = \emptyset$. The total weighted connection between the two sets is computed from the similarities encoded in the affinity matrix $\mathbf{U}$ and is defined as $\mathrm{links}(A, B) = \sum_{i\in A}\sum_{j\in B} U_{ij}$. It is easy to prove that the following identity holds true: $\mathrm{links}\,(A, V) = \mathrm{links}\,(A, B) + \mathrm{links}\,(A, A)$. Two measures of importance in graph partitioning are the cut, $\mathrm{cut}(A, B) = \mathrm{links}(A, B) = \sum_{i\in A}\sum_{j\in B} U_{ij}$, and the association, $\mathrm{assoc}(A, X) = \mathrm{links}(A, X) = \sum_{i\in A}\sum_{j\in X} U_{ij}$ where X stands for A, B or V. The graph cut is the central measure in graph partitioning problems. It represents the sum of the edge weights coming from