Neural Networks 59 (2014) 1–15

Contents lists available at ScienceDirect

Neural Networks

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

Learning Markov random walks for robust subspace clustering and estimation

Risheng Liu [a,*], Zhouchen Lin [b], Zhixun Su [a]

[a] *Dalian University of Technology, Dalian, China*

[b] *Key Lab. of Machine Perception (MOE), School of EECS, Peking University, Beijing, China*

ARTICLE INFO

Article history:
Received 15 July 2013
Received in revised form 18 April 2014
Accepted 11 June 2014
Available online 25 June 2014

Keywords:
Spectral clustering
Dimensionality reduction
Markov random walks
Transition probability learning
Subspace clustering and estimation

ABSTRACT

Markov Random Walks (MRW) has proven to be an effective way to understand spectral clustering and embedding. However, due to less global structural measure, conventional MRW (e.g., the Gaussian kernel MRW) cannot be applied to handle data points drawn from a mixture of subspaces. In this paper, we introduce a regularized MRW learning model, using a low-rank penalty to constrain the global subspace structure, for subspace clustering and estimation. In our framework, both the local pairwise similarity and the global subspace structure can be learnt from the transition probabilities of MRW. We prove that under some suitable conditions, our proposed local/global criteria can exactly capture the multiple subspace structure and learn a low-dimensional embedding for the data, in which giving the true segmentation of subspaces. To improve robustness in real situations, we also propose an extension of the MRW learning model based on integrating transition matrix learning and error correction in a unified framework. Experimental results on both synthetic data and real applications demonstrate that our proposed MRW learning model and its robust extension outperform the state-of-the-art subspace clustering methods.

1. Introduction

The graph spectral techniques (Chung, 1997; Von Luxburg, 2007) have many applications in machine learning, exploratory data analysis, computer vision and pattern recognition. Normalized Cut (NCut) (Shi & Malik, 2000), as one of the most successful spectral clustering methods, views the data set as a graph, whose nodes represent data points and whose edges are weighted according to the similarity between data samples. The success of such algorithms heavily depends on the choice of the affinity matrix. In addition to the graph cut interpretations, spectral clustering can also be understood in a probabilistic manner. The work in Meila and Shi (2001) views the local pairwise similarities as edge flows in Markov Random Walks (MRW) and studies the properties of the resulting transition matrix. In this view, the NCut criterion can be nicely interpreted in a general MRW framework. Along this direction, the work in Qiu and Hancock (2007) uses the commute time of a random walk for clustering and embedding. MRW can also be considered as a metric or a similarity structure over the data space, which is used by the clustering (Nadler, Lafon, & Coifman, 2005; Ng, Jordan, & Weiss, 2001) and embedding (Lafon & Lee, 2006) algorithms.

The most important problem with conventional MRW methods (e.g., Gaussian kernel MRW) is that they only consider the local similarity of the data set and there is no global structure constraint for the data. Thus these methods might be unsuitable for modeling data sampled from a mixture subspaces. The main reason is that the affinity in typical spectral methods is modeled only based on a characterization of "locality", which may fail to reveal the global subspace structure. However, in real applications several types of visual data, such as motion (Rao, Tron, Vidal, & Ma, 2010), face (Geng, Smith-Miles, Zhou, & Wang, 2011; Huang, Liu, & Metaxas, 2011) and video sequences (Mei & Ling, 2011; Wang, Tieu, & Grimson, 2010), have been known to be well characterized by subspaces. Therefore, there is a need to extend conventional spectral methods to model a mixture of subspaces. Recent advances in low-rank modeling have led to increasingly concise descriptions of the subspace structure. For instance, the work in Candès, Li, Ma, and Wright (2011) showed that the data points sampled from a single subspace can be exactly recovered by the rank minimization model. It is also shown in Liu, Lin, and Yu (2010) that the multiple subspace structure can be revealed by the "lowest rank" representation coefficients of a given dictionary. However, as discussed below, the spectrum properties of such representation matrix cannot be guaranteed. Thus further efforts

* Corresponding author. Tel.: +86 411 84708351.
E-mail addresses: rsliu0705@gmail.com, rsliu@dlut.edu.cn (R. Liu), zlin@pku.edu.cn (Z. Lin), zxsu@dlut.edu.cn (Z. Su).

http://dx.doi.org/10.1016/j.neunet.2014.06.005
0893-6080/

should be made to build upon the connection between the learnt representation and the affinity matrix used in spectral methods.

In this paper, by considering both the local pairwise similarity and the global subspace structure at the same time, we provide a new spectral framework from the MRW viewpoint for subspace clustering and estimation. Specifically, we learn a transition matrix by our local/global criteria and estimate a low-dimensional embedding from this graph. Then data points can be clustered into different subspaces in this feature space. In the following, we will review previous work on subspace clustering and then highlight the contributions of our research.

1.1. Previous work

A number of approaches for subspace clustering have been proposed in the past two decades. According to the mechanisms for data structure modeling, the existing works can be roughly divided into four main categories: algebraic, statistical, factorization, and compressive sensing methods.

Generalized Principal Component Analysis (GPCA) (Vidal, Ma, & Sastry, 2005) is an algebraic method for subspace clustering. The idea behind GPCA is that one can fit the data with polynomials. By representing subspaces with a set of homogeneous polynomials, subspace clustering is reduced to a problem of fitting data points with polynomials. This method does not impose any restriction on the subspaces. But the main drawback of GPCA is that it is difficult to estimate the polynomial coefficients when the data contains large noise. Recently, Robust Algebraic Segmentation (RAS) (Rao, Yang, Sastry, & Ma, 2010) has been proposed to resolve the robustness issue of GPCA. However, due to the computation difficulty in fitting large scale polynomials, RAS can only work for data with low-dimensionality and a small number of subspaces.

Statistical approaches usually model mixed data as a set of independent samples drawn from a mixture of probabilistic distributions (e.g., mixture of Gaussian). Then the problem of clustering is converted to a model estimation problem, which can be tackled by either Expectation–Maximization (EM) (Gruber & Weiss, 2004) or estimating the mixture structure by iteratively finding a min–max estimation (Fischler & Bolles, 1981). The Bayesian Ying-Yang harmony learning technique presented in Xu (0000) and Xu (2002) is a unified statistical framework to model unsupervised learning and recent investigations in Shi, Liu, Tu, and Xu (2014) show that this theory can be successfully applied for cluster number selection and determining the dimension for principal subspace. The main limitation of statistical models is the optimization difficulty. For example, due to the usage of EM algorithm, most statistical methods can only converge to a local minimum, thus are sensitive to initialization. Also, the sensibility to large errors and outliers is also a bottleneck for these methods.

The idea behind factorization methods (Costeira & Kanade, 1998; Gear, 1998) is to seek clustering from the factorization of the data matrix. The factorization can be computed from SVD (Costeira & Kanade, 1998) or the row echelon canonical form Gear (1998). However, all these methods are sensitive to noise. The work in Gruber and Weiss (2004) adds extra regularization terms to the formulation to reduce the effects of noise. Due to the optimization difficulty of the modified non-convex problem, this method may also get stuck at local minimum.

Compressive sensing has proven to be an extremely powerful tool for signal processing. Recently, there has been a surge of methods (Elhamifar & Vidal, 2009; Favaro, Vidal, & Ravichandran, 2011; Liu, Lin, De la Torre, & Su, 2012; Liu et al., 2010; Nasihatkon & Hartley, 2011; Ni, Sun, Yuan, Yan, & Cheong, 2010; Yu & Schuurmans, 2011) exploiting the discriminative nature of compact representation for subspace clustering. One type of methods, such as Sparse Subspace Clustering (SSC) (Elhamifar & Vidal, 2009; Nasihatkon & Hartley, 2011), is based on discovering the sparsest representations (SR) for the data set. According to the theoretical work of Nasihatkon and Hartley (2011), the within-subspace connectivity assumption for SSC holds only for 2- and 3-dimensional subspaces. In this view, it is possible for SSC to over-segment subspaces for dimension higher than 3. Therefore, extra post-processing stage is needed to overcome this intrinsic drawback for high dimensional data set.

Another type of method, such as Low-Rank Representation (LRR) (Liu et al., 2012, 2013, 2010), is based on minimizing the rank of the representation matrix. It has been proven that, under certain conditions, such non-convex problem can be efficiently solved by minimizing the nuclear norm (as a measure of 2D sparsity) of the matrix (Cai, Candès, & Shen, 2010). Theoretical analysis in Wei and Lin (0000) shows that in essence LRR is a kind of factorization method. Several extensions of this work have been developed. In Favaro et al. (2011), Favaro et al. extend the standard LRR to learn both clean dictionary and low-rank representation for subspace clustering. Indeed, a particular case of this method is equivalent to PCA (Jolliffe, 2002). Thus this method can also be utilized for single subspace estimation. A major drawback of this model is that it may be sensitive to sparse outliers due to the Frobenius norm measure for the noise term. The work in Yu and Schuurmans (2011) also proposes some theoretical analysis on LRR related optimization problems and proves that under the Simultaneous Block (SB) and/or Simultaneous Diagonal (SD) assumptions, a class of rank/norm based subspace clustering models can be solved in closed forms. However, due to the strict SB and SD assumptions on the data matrix, it is unclear whether or not their results can be extended to general problems and applied to real applications.

Overall, although compressive sensing based methods (i.e., SSC, LRR and their variations discussed above) all aim to learn an affinity matrix for spectral clustering, the spectrum properties (i.e., symmetric and nonnegative) of the representation matrix has been bypassed. Without consideration in this aspect, the validity of the constructed graph is poorly justified.

1.2. Our contribution

In this paper, we propose a novel method, called Low-Rank Markov Random Walks (LR-MRW), to learn a specific transition matrix (with low-rank property) to transfer the multiple subspaces structure from the observed data space to a low-dimensional discriminant feature space for subspaces clustering and estimation. Our motivations in this work are two-fold: the success of MRW in understanding spectral clustering and the matrix rank viewpoint for measuring the subspace structure.

On one hand, the intuition motivating this study is that since random walks reflect the combined effect of all possible weighted paths between a pair of nodes, the transition matrix can lead to a measure of cluster cohesion that is less sensitive than using edge weight alone, which underpins algorithms such as NCut. Therefore, it is natural to assume transition probabilities as a metric or a similarity measure over the data space for clustering. On the other hand, inspired by recent works on low-rank modeling (Liu et al., 2010; Wright, Ganesh, Rao, & Ma, 2009), we utilize rank as a measure of subspace structure for the transition matrix. In general, by introducing such local/global criteria, our work learns specific transition probabilities from the original data set to characterize both local pairwise relationship and global multiple linear subspaces structure. For noisy and corrupted data, we propose a robust extension of LR-MRW, which integrates transition matrix learning and noise corruption in a unified framework. Moreover, as a nontrivial byproduct, we propose closed-form solutions for a general class of nuclear norm regularized lease square problems. In the following, we highlight main contributions of the proposed approach: