Pattern Recognition Letters 68 (2015) 118–125

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Nonlinear subspace clustering using curvature constrained distances☆

Amir Babaeian[a,*], Mohammadreaza Babaee[b], Alireza Bayestehtashk[c], Mojtaba Bandarabadi[d]

[a] University of California San Diego, San Diego, CA USA
[b] Technische Universität München, Munich, Germany
[c] Oregon Health & Science University, Portland, OR, USA
[d] University of Coimbra, Coimbra, Portugal

ARTICLE INFO

Article history:
Received 2 February 2015
Available online 25 September 2015

Keywords:
Subspace clustering
Manifold
Isomap
Constrained shortest-path

ABSTRACT

The massive amount of high-dimensional data in science and engineering demands new trends in data analysis. Subspace techniques have shown remarkable success in numerous problems in computer vision and data mining, where the goal is to recover the low-dimensional structure of data in an ambient space. Traditional subspace methods like PCA and ICA assume that the data is coming from a single manifold. However, the data might come from several (possibly intersected) manifolds (surfaces). This has caused the development of new nonlinear techniques to cluster subspaces of high-dimensional data. In this paper, we propose a new algorithm for subspace clustering of data, where the data consists of several possibly intersected manifolds. To this end, we first propose a curvature constraint to find the shortest path between data points and then use it in Isomap for subspace learning. The algorithm chooses several landmark nodes at random and then checks whether there is a curvature constrained path between each landmark node and all other nodes in the neighborhood graph. It builds a binary feature vector for each point where each entry represents the connectivity of that point to a particular landmark. Then the binary feature vectors could be used as a input of conventional clustering algorithms such as hierarchical clustering. The performed experiments on both synthetic and real data sets confirm the performance of our algorithm.

1. Introduction

The amount of collected data has been increasing exponentially over the last decade. In many areas of machine learning, computer vision and data analysis, the data is represented by very high-dimensional features. For instance, images and videos are represented by millions of pixels. Therefore, the computational complexity of the data is a dramatic challenge in data processing which is referred to the "*curse of dimensionality*". However, high-dimensional data in most cases comes from low-dimensional structures instead of been uniformly distributed in ambient space [40]. Hence, many techniques have been proposed recently to recover the low-dimensional structure of the data from ambient space, the so-called subspace learning. Some work assume that the data is coming from a single structure like manifold learning techniques [40]. On the other hand, some methods assume that the data is coming from several structures.

In this paper, we propose a novel technique in recovering a low-dimensional representation of the data coming from several (possibly) intersected structures (surfaces). Therefore, the problem is multi-manifold clustering, which aims to label each data point according to the surface it comes from. This problem may exist in a number of applications, such as the extraction of galaxy clusters [24], road tracking [14], and target tracking [2–5,28]. For instance, in motion segmentation [13,22,36] and face recognition [6,12,20], the underlying surfaces are usually assumed to be linear or affine. In our approach, the main assumption is that the surfaces are smooth. In Fig. 1, the input and output of our algorithm is depicted. Here, we assume that the data is coming from three smooth manifolds (left) and the goal is to distinguish them with different labels/colors (left).

Several techniques have been proposed for multiple manifold clustering. However, most methods are designed for the case where manifolds do not intersect [11,25] or the manifolds that intersect have different either intrinsic dimensions or densities [1,16]. Basically, there are a few approaches aiming to recover the intersected manifolds. For instance, Souvenir et al. [31] implement a variant of K-means [7,8,32] where the cluster centers are considered as manifolds. Guo et al. [19] propose to minimize a (combinatorial) energy that includes local orientation information by using a tabu search. Recently, the state-of-the-art methods are based on local Principal Component Analysis (PCA). For example, the multi-scale spectral method of [21] uses the clustering routine of [17], which is developed in the

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

* Corresponding author. Tel.: +18582287952.
E-mail address: ababaeian@ucsd.edu, amir.babaeian@gmail.com (A. Babaeian).

http://dx.doi.org/10.1016/j.patrec.2015.09.001
0167-8655/

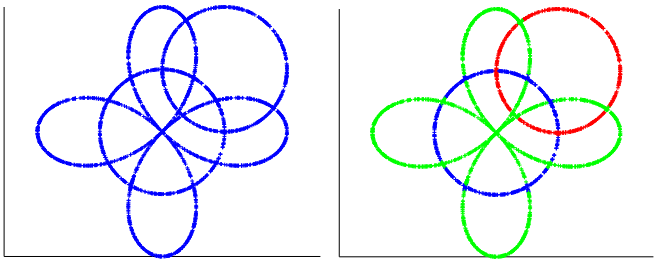

Fig. 1. Simulated data illustrating the problem of multi-manifold clustering. Left: Input data; Right: Output data obtained by our method.

context of semi-supervised learning and is inspired by the works of [38] and [18].

We propose a remarkably different approach for the problem of clustering multiple intersected manifolds based on connecting points to landmarks via curvature-constrained paths. Our approach can be interpreted as a constrained variant of [33], where we utilize a constrained shortest path distances. Isomap has been specifically designed for dimensionality reduction (or manifold learning) in a single-manifold setting used in different applications [26,35,39]. However, in particular, it cannot handle a self intersecting manifold . In our approach, the curvature constraint on paths prevents connecting points from one cluster to points from a different intersecting cluster. The algorithm is implemented as a simple variation of Dijkstra's algorithm.

The rest of the paper is organized as follows. In Section 2, we provide the related work in the area of multiple manifold clustering. Section 3 explains the notion of curvature constrained shortest-path and it's connection to the curvature constrained shortest-path. In Section 4, we introduce our algorithm for multi-manifold clustering. In Section 5, we provide the detailed information about the performed experiments on both synthetic and real data sets. There we compare the performance of our algorithm with several other algorithm. In addition, we discuss the robustness of our method to noise. We conclude our paper and provide an outline for future work in Section 6.

2. Related works

The last decade saw a flurry of propositions aiming at high-dimensional data clustering when the underlying clusters are not convex and particularly, in the situation where the points are sampled near low-dimensional structures. In the previous section, a few related works were introduced and now in this section, we elaborate on three of them, namely K-Manifolds (KM) [31], Spectral Curvature Clustering (SCC) [10], and Spectral Multi-Manifold Clustering (SMMC) [38]. Furthermore, we use them as benchmarks in our experiments. Our choice was dictated by performance, code availability and relevance to our particular setting.

The method of [21] renders impressive results but is hard to tune and relies on many parameters. The method of [18] is very similar to that of [38] and the code was not publicly available at the moment of writing this paper. The other methods for multi-manifold clustering (to the best of our knowledge) were not designed to resolve intersections of clusters of possibly identical intrinsic dimensions and sampling densities. Therefore, we chose the subspace clustering method of [10] among a few others methods that perform well in this context.

2.1. K-Manifolds

Souvenir et al. [31] suggest an algorithm that mimics K-means by replacing centroid points with centroid sub-manifolds. The method starts like Isomap by building a neighborhood graph and computing shortest path distances within the graph. After randomly initializing a K-by-n weight matrix, $W = (w_{ki})$, where w_{ki} represents the probability that point i belongs to the kth cluster, it alternates between an M-Step and an E-Step. In the M-Step, for each k, the points are embedded in $\mathbb{R}^K$ using a weighted variant of multidimensional scaling using the weights $(w_{ki} : i = 1, \ldots, n)$. In the E-Step, for each k and i, the normal distance of point x_i to the cluster k is estimated as

$$\delta_{ki} = \frac{\sum_j w_{kj}(d(x_i, x_j) - d_k(x_i, x_j))}{\sum_j w_{kj}},$$

where $d(x_i, x_j)$ denotes the shortest path distance in the neighborhood graph and $d_k(x_i, x_j)$ denotes the Euclidean distance in the kth embedding, between points x_i and x_j. The weights are then updated as $w_{ki} \propto \exp(-d_{ki}^2/\sigma^2)$ such that $\sum_k w_{ki} = 1$ for all i, where σ^2 is chosen automatically.

2.2. Spectral curvature clustering

Chen et al. [10] propose a spectral method for subspace clustering based on the assumption that the underlying surfaces are affine. However, we compare our method to theirs when the surfaces are affine and also when the surfaces are curved. The latter is done as a proof of concept, whereas it is clear that this method cannot handle curved surfaces, like any other method for subspace clustering. The procedure assumes that all subspaces are of the same dimension d, which is a parameter of the method. For each $(d+2)$-tuple, $x_{i_1}, \ldots, x_{i_{d+2}}$, it computes a notion of curvature $C_{i_1,\ldots,i_{d+2}}$ which measure how well this $(d+2)$-tuple is approximated by an affine subspace of dimension d. After reducing the tensor $\mathbf{C} = (C_{i_1,\ldots,i_{d+2}} : i_t = 1, \ldots, N)$ spectral graph partitioning [25] is applied.

2.3. Spectral multi-manifold clustering

Wang et al. [38] propose a spectral clustering method using a dissimilarity function that factors in the Euclidean distance and the discrepancy between the local orientation of the data points. The surfaces are assumed to be of the same dimension d and this number should be known as a prior. First, a mixture of probabilistic principal component analyzers [34] are fitted to the data, approximating the point cloud by a union of d-planes. This is used to estimate the tangent subspace at each data point. The dissimilarity between two data points is then an increasing function of their Euclidean distance and also the principal angles between their respective affine subspaces. These dissimilarities are fed into the spectral graph partitioning method proposed by [25].

2.4. Spectral clustering

Spectral clustering has been widely used as a clustering method for high-dimensional data. First, it forms a similarity matrix of the points where elements of this matrix are the relative similarity between each pair of points in a high-dimensional space. Then it computes the normalized or unnormalized Laplacian matrix L. By computing the first K eigenvectors of L we form a new matrix U, where the number of columns of U is equal to K. The rows of U are the low-dimensional representation of high-dimensional data. By applying the k-means algorithm to the rows of U we extract the cluster membership for each point. Spectral clustering is closely related to the nonlinear dimensionality reduction methods like locally-linear embedding.