Neurocomputing 143 (2014) 109–122

Contents lists available at ScienceDirect

Neurocomputing

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

A novel graph-based *k*-means for nonlinear manifold clustering and representative selection

Enmei Tu [a], Longbing Cao [b], Jie Yang [a,*], Nicola Kasabov [c]

[a] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, China
[b] UTS Advanced Analytics Institute, University of Technology Sydney, Australia
[c] The Knowledge Engineering and Discovery Research Institute, Auckland University of Technology, New Zealand

ARTICLE INFO

Article history:
Received 6 November 2013
Received in revised form
4 April 2014
Accepted 20 May 2014
Communicated by Liang Wang
Available online 20 June 2014

Keywords:
k-means
Manifold clustering
Random walk
Graph learning

ABSTRACT

Many real-world applications expose the nonlinear manifold structure of the lower dimension rather than its high-dimensional input space. This greatly challenges most existing clustering and representative selection algorithms which do not take the manifold characteristics into consideration. The performance of the corresponding learning algorithms can be greatly improved if the manifold structure is considered. In this paper, we propose a graph-based *k*-means algorithm, GKM, which bears the simplicity of classic *k*-means while incorporating global information of data geometric distribution. GKM fully exploits the intrinsic manifold structure for appropriate data clustering and representative selection. GKM is evaluated on both synthetic and real-life data sets and achieves very impressive results compared to the state-of-the-art approaches, including classic *k*-means, kernel *k*-means, spectral clustering, and clustering through ranking and for representative selection. Given the widespread appearance of manifold structures in real world problems, GKM shows promising potential for partitioning manifold-distributed data.

1. Introduction

Clustering is aimed to divide a set of samples $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}$ into K disjointed subsets $\mathcal{X}_1,\mathcal{X}_2,\ldots,\mathcal{X}_K$ so that points in the same subset share common properties while points which belong to different subsets do not share these properties. There are many algorithms available for performing this task, with *k*-means [1] being the most popular because of its properties as follows: it can effectively partition data with Gaussian-like distributions; it is intuitive and easy to implement; its implementation is of linear computational complexity, etc. However, *k*-means also exhibits typical disadvantages for manifold distributed data sets and our motivation for extending *k*-means to a manifold algorithm is based on the following observations. (1) Recently, it was widely acknowledged in the data analysis and machine learning communities that many real world data sets, such as face images, voice spectrum, hand-writing digital images and document contents, stringently obey the low-rank rules, which means that they are distributed on a manifold whose dimensionality is often much lower than the ambient space [2,3]. The classic *k*-means algorithm, CKM, does not take this important characteristic into account and thus performs poorly when dealing with such data sets. (2) The cluster centers (or prototypes) given by CKM are generally not members of a data set and thus cannot be used directly for many applications, such as video and text summarization [4,5], which aim to choose a small subset of frames or sentences that best describe the overall contents. Compared to CKM, the kernel *k*-means [6], KKM, in most cases, yields better results by mapping samples to a possibly much lower dimensional space than that in the input space, in which the manifolds of different classes are separable. Unfortunately, such a perfect mapping may not exist in practice, nor is it clear what kinds of mapping exist for a given data set.

Inspired by the isometric feature mapping (isomap)-based nonlinear dimension reduction algorithm [3] and the manifold ranking algorithm [7], we propose a graph-based *k*-means algorithm, GKM, to overcome the above disadvantages by taking the geometric characteristics of data distribution into account. Particularly, GKM can fully exploit the underlying manifold structure of a data set to produce better clustering results and, meanwhile, identify a suitable data point representative (or centroid) for each subset. GKM also retains the advantages of CKM, such as ease of implementation, intuitive and low computational complexity. Extensive experiments conducted on both synthetic and real-world data sets validate that GKM is very effective for manifold clustering with appropriate representative selection.

* Corresponding author.
E-mail address: jieyang@sjtu.edu.cn (J. Yang).

http://dx.doi.org/10.1016/j.neucom.2014.05.067
0925-2312/

The remainder of this paper is organized as follows: Section 2 reviews the related works in nonlinear manifolds clustering and representative selection and Section 3 briefly reviews the classic k-means algorithm. Section 4 presents our graph-based k-means algorithm. Finally, the simulations and comparisons are presented in Section 5, followed by discussions and conclusions in Section 6.

2. Related works

Nonlinear manifold clustering and representative selection are very challenging topics in machine learning. The work in [8,9] extends mean shift to nonlinear manifold clustering and performs well in motion segmentation, but it is restricted to analytic manifolds. Goh and Vidal [10] only consider separated nonlinear manifold clustering. In recent years, spectral clustering [11–15] and manifold clustering algorithms [16–18] have been proposed to handle general manifold-distributed data sets. These approaches either do not present a meaningful subset of data points which can mostly represent the data set or need a complicated optimization process. Other algorithms have been specially designed for the selection of representative data points [5,4] on the condition of solving an unsmooth optimization problem, thus suffering from high computational costs in dealing with large data sets. In addition, they also do not present a clear description of data clusters and thus cannot tell which points are best represented by which representative. For other similar data-driven researches and applications, we recommend the recent results in [19,20]. Very limited work has been conducted on k-means for manifolds, and as yet, this area has not been fully exploited. The work in [21] is restricted to data on sphere and [22] aims to analyze the reconstruction error and learning rate of k-means on manifolds but does not provide a concrete algorithm for clustering.

There are also studies of k-means on graph. In [23] Euclidean distance and centroid are replaced by graph edit distance and the so-called mean graph respectively, but the computational cost of computing both graph edit distance and mean graph are very high and thus make the algorithm not suitable for large data sets. In [24] the classic k-means is just utilized as a post-processing method after thresholding the sequence of edge lengths that added to the minimal spanning tree by Prims algorithm to obtain the final clusters.

In contrast, in this paper we make two essential changes to the classic k-means for dealing with nonlinear manifold data. Particularly, we first extend the centroid concept of point cloud in Euclidean geometry to the centroid of manifold in Riemannian geometry. Then, borrowing from graph-based semi-supervised learning method, a new random walk model, the tired random walk model which is capable of describing the similarity between points on nonlinear manifold, is proposed to determine the centroid–member relationships on graph.

3. Review of the classic k-means clustering algorithm

Let $\mathcal{X}=\{x_1,x_2,\ldots,x_n\}\in\mathbb{R}^d$ be a sample set to be partitioned into K groups and $\mathcal{Y}=(y_1,y_2,\ldots,y_n)$, $y_i\in\{1,2,\ldots,K\}$ be the label vector in which each component gives the class label of the corresponding sample in $\mathcal{X}$. Clustering is aimed to partition $\mathcal{X}$ into K disjointed subsets $\mathcal{X}_1,\mathcal{X}_2,\ldots,\mathcal{X}_K$ so that samples in the same subset have same class labels, i.e. $y_i=k$, $x_i\in\mathcal{X}_k$, $k\in\{1,\ldots,K\}$. We denote by $\mathcal{C}_k=\{i_1,i_2,\ldots,i_{|\mathcal{C}_k|}\}$ the index set of elements in subset $\mathcal{X}_k$.

In classic k-means (CKM), the algorithm updates iteratively the cluster centroids $\{c_1,c_2,\ldots,c_K\}$ and index sets $\mathcal{C}_k, k=1..K$. For a particular cluster, the centroid of the cluster is updated by simply averaging the memberships over its members, and the membership of a sample is determined by the nearest Euclidean distance from it to all the centroids. Mathematically, the algorithm updates the cluster centroid by

$$c_k=\frac{1}{|\mathcal{C}_k|}\sum_{i\in\mathcal{C}_k}x_i \tag{1}$$

After all centroids are updated, the algorithm re-computes the Euclidean distance $d_E(x,y)=\|x-y\|^2$ between each sample and all the cluster centroids and labels the sample as a member of the cluster whose centroid achieves the smallest Euclidean distance

$$y_i=\arg\min_{k=1\ldots K} d_E(x_i,c_k) \tag{2}$$

In the first iteration, the algorithm randomly chooses K points $\{c_1,c_2,\ldots,c_K\}$ to be the initial centroids. Thereafter, these two steps iterate alternatively until the algorithm converges.

4. A graph k-means manifold clustering algorithm

In this section, we propose a graph-based k-means algorithm, GKM, which takes the intrinsic manifold structure into account. Similar to classic k-means, GKM also has two essential steps, updating centroids and updating the samples membership, but these steps differ in nature from their counterparts in the classic k-means. We also present an initialization method to obtain a group of high quality initial centroids. These are described in detail as follows.

4.1. Updating centroids

In the classic k-means, the k-th centroid obtained by Eq. (1) is the coordinates mean of the point cloud. This is the classic centroid concept in Euclidean geometry. Here we extend this concept to Riemannian geometry, *i.e.* computing the centroid of a manifold. Note that the centroid in Eq. (1) is actually the optimal solution of the following optimization problem:

$$c_k=\arg\min_{x\in\mathbb{R}^d}\frac{1}{2}\sum_{i\in\mathcal{C}_k}d_E(x,x_i),\quad k=1\ldots K \tag{3}$$

where $d_E(x,x_i)=\|x-x_i\|^2$ is the Euclidean distance. For data points sampled from manifolds, this optimization has two defects: (a) the centroid may move off the manifold and thus it cannot represent the underlying data distribution well; (b) the Euclidean distance cannot reflect the true relationships between samples in Riemannian geometry because it does not fully capture the data geometric feature and thus it is not a proper measure of similarity.

To capture the intrinsic geometric feature, we generalize problem (3) by (a) restricting the centroid on the manifold[1]; and (b) using geodesic distance as a measurement between two points. Combining these two we extend the classic centroid in Euclidean geometry to the manifold centroid in Riemannian geometry by solving the following optimization problem:

$$c_k=\arg\min_{x_j,j\in\mathcal{C}_k}\frac{1}{2}\sum_{i\in\mathcal{C}_k}d_g(x_j,x_i),\quad k=1\ldots K \tag{4}$$

where $d_g(x_i,x_j)$ is the geodesic distance between two samples x_i and x_j. However, in clustering settings, the exact geodesic distance between two samples x_i and x_j usually cannot be obtained directly, because we have no prior information about the underlying

[1] To restrict centroid on manifold, it is also possible to adopt other method to choose the class centroid, such as the medium point. But our optimization method has at least two benefits: first, the formulation is straightway to classic k-means practitioners and easy to be understood; second, it is of computational efficiency, as will be demonstrated.