Neurocomputing ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Neurocomputing

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

Joint diversity regularization and graph regularization for multiple kernel k-means clustering via latent variables

Teng Li*, Yong Dou, Xinwang Liu

National Laboratory for Parallel and Distributed Processing, National University of Defense Technology, Changsha 410073, PR China

ARTICLE INFO

Article history:
Received 2 February 2016
Received in revised form
23 June 2016
Accepted 16 August 2016
Communicated by: Feiping Nie

Keywords:
Clustering
Kernel k-means
Latent variables
Diversity regularization
Graph regularization

ABSTRACT

Multiple kernel k-means (MKKM) clustering algorithm is widely used in many machine learning and computer vision tasks. This algorithm improves clustering performance by extending the traditional kernel k-means (KKM) clustering algorithm to a multiple setting by combining a group of pre-specified kernels. In this paper, we develop and propose a multiple kernel k-means clustering via latent variables (MKKLV) algorithm, in which base kernels can be adaptively adjusted with respect to each sample. To improve the effectiveness of the kernel-specific and sample-specific characteristics of the data, joint diversity regularization and graph regularization are utilized in the MKKLV algorithm. An efficient three-step iterative algorithm is employed to jointly optimize the kernel-specific and sample-specific coefficients. Experiments validate that our algorithm outperforms state-of-the-art techniques on several different benchmark datasets.

1. Introduction

In many machine learning [1], computer vision [2,3], and data mining [4,5] fields, clustering algorithms are used to find meaningful groupings of samples in an unsupervised manner. One example of traditional clustering algorithms is k-means clustering [6–9], which uses the k prototype to characterize the data and minimizes the sum-of-squares cost function. However, the standard k-means algorithm is limited to the sum-of-squares cost function, and it cannot identify arbitrarily shaped clusters. An advantage of kernel-based clustering methods, such as kernel k-means clustering [10,11], is its capability to handle non-linear separable clusters, typically with good clustering performance. In a practical scenario, different kernels can be constructed because samples have multiple representations that originate from different data sources (i.e. different kernel functions can lead to different kernels; different kernels can also be constructed through multiple feature representations). Multiple kernel k-means clustering is superior over other approaches because it utilizes all of the available information for performing kernel k-means, which is better than using a single fixed kernel.

Multiple kernel k-means clustering has been attracting increasing attention in recent years, and many efforts have been made to improve clustering performance [12–16]. [12] proposes a multiple view clustering algorithm that incorporates multiple kernels and automatically adjusts kernel weights. In [13], they assign kernel weights to the corresponding view's information and utilize a parameter to control the sparsity of these weights. [14] combines the kernels calculated on the views in a localized manner to better capture the sample-specific characteristics of the data. In [15], they propose a localized multiple kernel clustering method, which is dedicated to the dataset with varying local distributions. [16] presents a robust multiple kernel k-means algorithm by replacing the sum-of-squared loss with a $\ell_{2,1}$– norm.

Existing state-of-the-art multiple kernel k-means clustering methods are hindered by several defects. First, [12,13,16] learn the kernel combination weights on the basis of a given dataset and use the same kernel weights for all samples, however, such an approach may not be ideal because of sample-specific characteristics of the data. [14] learns the sample-specific combination weights directly instead of the kernel combination weights, however, this method cannot capture kernel-specific information. An ideal multiple kernel k-means clustering algorithm should consider both the kernel-specific and sample-specific information, which leads to a fairer learning model between the kernel-specific and sample-specific properties. This mechanism is lacking in existing multiple kernel k-means clustering methods.

Second, the data fitting term is very important for multiple kernel k-means clustering because the coefficients are learned from minimizing the regularized fitting error. Good fitting terms in multiple kernel k-means clustering should satisfy two properties: (1) enable the fitting model to capture more information among

* Corresponding author.
E-mail addresses: liteng09@nudt.edu.cn (T. Li), yongdou@nudt.edu.cn (Y. Dou), 1022xinwang.liu@gmail.com (X. Liu).

http://dx.doi.org/10.1016/j.neucom.2016.08.060
0925-2312/

different kernels and different samples; (2) prevent the learning model from becoming over-flexible. However, in [14], the optimization problem contains many real-valued variables (number of kernels × number of samples), yet the optimization problem has no proper regularization term. Such behavior will result in an over-flexible learning model.

Basing on the observations and findings above, we propose a novel multiple kernel k-means clustering method, i.e. the multiple kernel k-means clustering via latent variables (MKKLV), where base kernels can be adaptively adjusted with respect to each sample. To the best of our knowledge, this study is the first to combine the kernel-specific and sample-specific information into a joint formulation, which considers both the kernel-specific and sample-specific information in improving clustering performance. To fully utilize the kernel information, we adopt a diversity regularization, which has been proven effective in [17], to capture the complementary information among different kernels. In addition, to avoid the over-flexible problem induced by latent variables, we adopt a graph regularization. As will be demonstrated in subsequent sections, integrating the diversity regularization and graph regularization terms into our MKKLV formulation allows the latent variables to be optimized and the kernel combination weights to be well separated. In this way, the proposed algorithm retains the advantages of existing efficient optimization algorithms.

In summary, we highlight the main contributions of this paper as follows:

- We propose a MKKLV algorithm by exploring the multiple kernel k-means with latent variables, in which base kernels can be adaptively adjusted with respect to each sample.
- To improve the effectiveness of the kernel-specific and sample-specific characteristics of the data, we further use a joint diversity regularization and graph regularization for the MKKLV algorithm.
- We derive and present an efficient three-step iterative algorithm to optimize the kernel-specific and sample-specific coefficients jointly.
- We conduct comprehensive experiments to compare the proposed approach with existing state-of-the-art methods on six benchmark datasets. The experimental results demonstrate the superiority of the proposed method over the state-of-the-art methods.

The remainder of this paper is organized as follows. Section 2 gives the notations and preliminaries used throughout the paper. We introduce the kernel k-means and multiple kernel k-means in Section 3. We then present the MKKLV algorithm in Section 4. An efficient iterative algorithm is proposed in Section 5, where the details of the algorithm are also provided. The convergence analysis and time complexity are presented in Section 6. We compare the clustering performance of MKKLV and state-of-the-arts multiple kernel k-means clustering algorithms, and discuss the parameter sensitivity in Section 7. Finally, conclusions are drawn in the Section 8.

2. Notations and preliminaries

Throughout the paper, matrices are written as boldface capital letters and vectors are denoted as boldface lowercase letters. For matrix $\boldsymbol{M}$, $tr(\boldsymbol{M})$ is the trace of $\boldsymbol{M}$ if $\boldsymbol{M}$ is square.

Assume that we have n data samples $\{x_i\}_{i=1}^n$, let $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]^T$ denote the data matrix with each row being a data feature vector, where $\boldsymbol{x_i} \in R^d$ is the feature descriptor of the i-th example.

In manifold learning, graph Laplacian is defined by $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, where W_{ij} is the edge weight between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ in the sparse adjacency matrix on the neighborhood graph (e.g. one can use Gaussian kernel or K-nearest neighbors) and $\boldsymbol{D}$ is a diagonal matrix with its i-th diagonal entry being $D_{ii} = \sum_j W_{ij}$.

3. Multiple kernel k-means

In this section, we first present the kernel k-means, which transforms the sum-of-squares minimization cost function of the traditional k-means into a trace minimization problem. Then we extend it to the multiple kernel k-means.

3.1. Kernel k-means

Suppose we are given a dataset $X = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]^T \in R^{n\times d}$, and aim to partition these data point into k disjoint clusters. The objective of k-means can be written as follows

$$\min \sum_i \left\|\boldsymbol{x_i} - \boldsymbol{\mu}_{z_i}\right\|^2 \quad \text{s.t. } z_i = 1, 2, \ldots k. \tag{1}$$

where we want to minimize over the following iterations

$$z_i = \arg\max_k \left\|\boldsymbol{x_i} - \boldsymbol{\mu_k}\right\|^2 \tag{2}$$

$$\boldsymbol{\mu_k} = \frac{1}{N_k}\sum_{i\in c_k} \boldsymbol{x_i} \tag{3}$$

Now we assume that we have defined a feature mapping $\Phi(\cdot)$ that maps the data samples into a feature space. We introduce a $n \times k$ assignment matrix, $\boldsymbol{Z} \in \{0, 1\}^{n\times k}$, each column of which represents a data-case and contains exactly one 1 at row k if it is assigned to cluster k. The objective of kernel k-means is to minimize the sum-of-squares cost function over the cluster assignment variables.

$$\min_{\boldsymbol{Z}\in\{0,1\}^{n\times k}} \sum_{i=1,c=1}^{n,k} z_{ic}\left\|\Phi(x_i) - \boldsymbol{\mu}_c\right\|_2^2 \quad \text{s.t. } \sum_{c=1}^{k} z_{ic} = 1, \forall\, i \tag{4}$$

where $\boldsymbol{\mu}_c = \frac{1}{n_c}\sum_{i=1}^n z_{ic}\Phi(x_i)$ is the centroid of cluster c ($1 \le c \le k$) and $n_c = \sum_{i=1}^n z_{ic}$.

Then we define $\boldsymbol{L} = \mathrm{diag}\left(n_1^{-1}, n_2^{-1}, \ldots, n_k^{-1}\right)$ and $\boldsymbol{\Phi} = [\Phi(x_1)\ \Phi(x_2)\ \ldots\ \Phi(x_n)]$, and the optimization equation in kernel k-means can be converted into an equivalent matrix-vector form problem

$$\min_{\boldsymbol{Z}\in\{0,1\}^{n\times k}} \mathrm{tr}((\boldsymbol{\Phi} - \boldsymbol{M})^T(\boldsymbol{\Phi} - \boldsymbol{M})) \quad \text{s.t. } \boldsymbol{Z1}_k = \boldsymbol{1}_n, \boldsymbol{M} = \boldsymbol{\Phi ZLZ}^T. \tag{5}$$

Next we can find that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{L}^{-1}$ and $(\boldsymbol{ZLZ}^T)^2 = \boldsymbol{ZLZ}^T$. Using this we can obtain the following equation

$$\mathrm{tr}((\boldsymbol{\Phi} - \boldsymbol{M})^T(\boldsymbol{\Phi} - \boldsymbol{M})) = \mathrm{tr}(\boldsymbol{K}) - \mathrm{tr}(\boldsymbol{L}^{1/2}\boldsymbol{Z}^T\boldsymbol{KZL}^{1/2}) \tag{6}$$

where $\boldsymbol{\Phi}^T\boldsymbol{\Phi} = \boldsymbol{K}$. Therefore the optimization problem can be formulated as follows

$$\min_{\boldsymbol{Z}\in\{0,1\}^{n\times k}} \mathrm{tr}(\boldsymbol{K}) - \mathrm{tr}(\boldsymbol{L}^{1/2}\boldsymbol{Z}^T\boldsymbol{KZL}^{1/2}) \quad \text{s.t. } \boldsymbol{Z1_k} = \boldsymbol{1_n}. \tag{7}$$

It should be noted that this problem is very difficult to solve due to the constraint is to search of discrete matrix $\boldsymbol{Z}$. However, this problem can be approximated through a relaxation on this constraint. Recall that $\boldsymbol{Z}^T\boldsymbol{Z} = \boldsymbol{L}^{-1}$, thus $(\boldsymbol{L}^{1/2}\boldsymbol{Z}^T\boldsymbol{ZL}^{1/2}) = \boldsymbol{I}_k$. By renaming $\boldsymbol{H} = \boldsymbol{ZL}^{1/2}$, we can formulate the following relaxation of the problem

Please cite this article as: T. Li, et al., Joint diversity regularization and graph regularization for multiple kernel k-means clustering via latent variables, Neurocomputing (2016), http://dx.doi.org/10.1016/j.neucom.2016.08.060