



Locality sensitive C-means clustering algorithms

Pengfei Huang, Daoqiang Zhang*

Department of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

ARTICLE INFO

Article history:

Received 25 December 2009

Received in revised form

22 July 2010

Accepted 29 July 2010

Communicated by X. Gao

Available online 18 September 2010

Keywords:

Locality sensitive weight

Fuzzy C-means (FCM)

Semi-supervised clustering

ABSTRACT

The concept of preserving locality information in dimensionality reduction and semi-supervised classification have been very popular recently. In this paper, we attempt to use locality sensitive weight for clustering, where the neighborhood structure information between objects are transformed into weights of objects. We develop two novel locality sensitive C-means algorithms, i.e. Locality-weighted Hard C-Means (LHCM) and Locality-weighted Fuzzy C-Means (LFCM), following the standard C-Means and fuzzy C-means, respectively. In addition, two semi-supervised extensions of LFCM are proposed to better use some given partial supervision information in data objects. Experimental results on both artificial and real datasets validate the effectiveness of the proposed algorithms.

© 2010 Elsevier B.V. All rights reserved.

1. Introduction

Clustering deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters. At present, clustering algorithms can be categorized into several types, such as partitional method, hierarchical method, density-based method, grid-based method and model-based method [1].

In this paper, we mainly focus on the partitional method. Presently, most clustering algorithms treat all data samples equally in the clustering process, such as hard C-Means (HCM) and its fuzzy extension, i.e. fuzzy C-Means (FCM) [2]. However, different samples may play different roles in the clustering process, because the samples distribute nonuniformly and asymmetrically. Moreover, a sample may contribute to the clustering results differently in different processes. Hence, it is very useful to give an appropriate sample weight in cluster analysis. For that purpose, sample weighting clustering algorithm have been proposed in literature [3–7].

In sample weighting clustering, the weight of each sample is very important, since it determines the impact of the sample on the clustering analysis. Conditional fuzzy C-means [3] and deterministic annealing clustering [4] consider various contributions of different samples and take account of sample weighting. However, the application of the above algorithms are limited because they need users or heuristic principle to weight samples.

To overcome that problem, Nock and Nissenslen proposed a formalized clustering framework, borrowing the idea of the boosting algorithm, which offers penalizing solutions via weights

on the samples [5]. In their paper [6], they pointed out the importance of calculating the sample weight automatically during the process of clustering analysis. Li et al. have proposed a typical-weighting clustering algorithm for large datasets. It can obtain original clustering samples using the atom-clustering algorithm, then weight them according to the atom number of samples [7]. Zhang et al. have introduced the document clustering algorithm based on sample weighting, which utilizes PageRank value as the weight of the samples and then assigns different weights to various samples, such that more reasonable centers could be obtained [8]. However, it is only applicable to document clustering and related areas. Gao et al. have presented weighted fuzzy C-means clustering, which considers the appearance probability of the gray levels from the gray histogram in an image as the weight parameter and hence improves the algorithm efficiency [9]. However, it is only suitable for image data. Recently, the weighting idea has also been used for clustering of fuzzy and relational data, respectively [10,11].

On the other hand, in machine learning and pattern recognition community, there have been a recent trend to utilize the local structural information for learning. For example, The concept of preserving locality information in dimensionality reduction and semi-supervised classification have been very popular recently [12,13]. Literatures [14–16] effectively utilizes the structure information by building a graph incorporating neighborhood information of the dataset. Using the notion of the graph Laplacian, a weight matrix which indicates the intrinsic structure is set up. However, to the best of knowledge, it remains unknown whether the local structure information among the clustering objects is also helpful to sample weighting clustering.

In this paper, motivated by the idea of optimally preserving the neighborhood structure in dimensionality reduction and semi-supervised learning, we propose a novel locality preserving

* Corresponding author.

E-mail address: dqzhang@nuaa.edu.cn (D. Zhang).

weighting scheme for clustering, from which two new algorithms, i.e. Locality-weighted Hard C-Means (LHCM) and Locality-weighted Fuzzy C-Means (LFCM) are developed. LHCM and LFCM calculate the distance between the samples and the centers to gain a proper weight parameter so that they can primarily describe the neighborhood structure of the data. In addition, the proposed methods are extended for semi-supervised cases to use the available supervision information in data, e.g. partial labeled data or pairwise constraints which specify whether a pair of data belong to the same class or not [16].

The rest of this paper is organized as follows: In Section 2 the background on HCM and FCM are briefly described. Section 3 derives the proposed LHCM and LFCM Clustering algorithms in detail. Section 4 gives the semi-supervised extensions on LFCM. The experimental results are given in Section 5. Finally, we conclude this paper in Section 6.

2. Background

2.1. HCM

HCM is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The algorithm classifies n vectors x_j ($j=1,2,\dots,n$) through a certain number of clusters (assume c clusters G_i ($i=1,2,\dots,c$) fixed a priori, and calculates each centroid v_i aiming at minimizing the objective function. The objective function is defined as follows:

$$J = \sum_{i=1}^c \sum_{x_j \in G_i} \|x_j - v_i\|^2 \quad (1)$$

2.2. FCM

FCM is a method of clustering which allows one piece of data to belong to two or more clusters. It is based on minimization of the following objective function:

$$J(U, v_1, \dots, v_c) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m \|x_j - v_i\|^2 \quad (2)$$

where x_j is the j th data example, v_i is the i th cluster center, and u_{ij} is the degree of membership of x_j in the cluster i . The weighting exponent m is a real number greater than 1 and the appropriate values depend on datasets. The theoretical analysis on the parameter m can be seen in the Refs. [17,18]. Finally in (2), $\|\cdot\|$ is a norm measuring the distance metric between data examples and the cluster centers. Fuzzy partitioning is carried out through an alternate iterative optimization [2,19] of the objective function shown above, with some properties:

$$\sum_{i=1}^c u_{ij} = 1, \quad \forall j = 1, \dots, n, \quad 0 < u_{ij} < 1, \quad 0 < \sum_{j=1}^n u_{ij} < n \quad (3)$$

3. Locality sensitive C-means clustering

3.1. Locality-weighted hard C-means (LHCM)

Suppose that $X = x_1, x_2, \dots, x_n$ is a d -dimensional database with n points, and is divided into c clusters $v = v_1, v_2, \dots, v_c$, each cluster can be represented by its cluster center v_i . The objective function of LHCM is defined as follows:

$$J = \sum_{i=1}^c \sum_{x_j \in G_i} s_{ij} \|x_j - v_i\|^2 \quad (4)$$

where G_i denotes the i th cluster and s_{ij} is the weight between points $\{x_j\}$ and centers $\{v_i\}$. To preserve the neighborhood structure information in the weight, we define the weighting function as follows:

$$s_{ij} = e^{-\|x_j - v_i\|^2 / t_i} \quad (5)$$

where t_i is a scaling parameter. When $t_i \rightarrow 0$, the weight matrix becomes the most important ingredient of the clustering result while the weights are very similar with each other. In this case, the weighted clustering will generate much poorer clustering result. On the other hand, when $t_i \rightarrow \infty$, the weight matrix has entries all equal to 1, and thus the weighted clustering is degraded into non-weighted clustering.

In order to choose appropriate values for the weights, we use a local scale for t_i as follows:

$$t_i = \begin{cases} \sigma_i^2 & x_j \in N_{ik} \\ \left(\frac{1}{c} \sum_{i=1}^c \sigma_i \right)^2 & \text{otherwise} \end{cases} \quad (6)$$

where $\sigma_i = (1/k) \sum_{j=1}^k \|x_j - v_i\|^2$, k is the number of the neighbors of the i th center. N_{ik} is the k Nearest Neighbor (k -NN) neighborhoods of the i th cluster. From (6), we can see that the scale can automatically adapt to the local structure. In practice, usually it is much easier to choose values for k than for t_i .

Let $u_{ij} \in \{0,1\}$ denote whether $x_j \in G_i$ or not, i.e. $u_{ij}=1$ means $x_j \in G_i$, and vice versa. Following the standard HCM, given the locality weight s_{ij} we can easily derive the solutions of LHCM by the following alternate iterations between the indicator u_{ij} and the cluster centers v_i . The detailed pseudo-code of LHCM is listed in Table 1.

$$u_{ij} = \begin{cases} 1 & \text{if } \forall k, \|x_j - v_i\|^2 \leq \|x_j - v_k\|^2 \\ 0 & \text{otherwise} \end{cases} \quad (7)$$

$$v_i = \frac{\sum_{j=1}^n u_{ij} s_{ij} x_j}{\sum_{j=1}^n u_{ij} s_{ij}} \quad (8)$$

3.2. Locality-weighted fuzzy C-Means (LFCM)

As in LHCM, we modify the standard FCM by introducing the locality weight s_{ij} . The objective function of LFCM is defined as follows:

$$J(U, v_1, \dots, v_c) = \sum_{i=1}^c \sum_{j=1}^n u_{ij}^m s_{ij} \|x_j - v_i\|^2 \quad (9)$$

where x_j is the j th of d -dimensional measured data, v_i is the i th cluster center, u_{ij} represents the fuzzy membership of the j -th point with respect to cluster i , s_{ij} is the locality weight between points x_j and centers v_i . The parameter m is a weighting exponent

Table 1

The LHCM algorithm.

Initialize: the cluster centers $v^{(0)} = \{v_1^{(0)}, v_2^{(0)}, \dots, v_c^{(0)}\}, l = 0, \epsilon > 0$
Step 1: update $s_{ij}^{(l+1)}$ by the equation: $s_{ij}^{(l+1)} = e^{-\ x_j - v_i^{(l)}\ ^2 / t_i}$
Step 2: update $u_{ij}^{(l+1)}$ with the equation: $u_{ij}^{(l+1)} = \begin{cases} 1 & \text{if } \forall k, \ x_j - v_i\ ^2 \leq \ x_j - v_k\ ^2 \\ 0 & \text{otherwise} \end{cases}$
Step 3: update $v_i^{(l+1)}$ with the equation: $v_i^{(l+1)} = \frac{\sum_{j=1}^n u_{ij}^{(l+1)} s_{ij}^{(l+1)} x_j}{\sum_{j=1}^n u_{ij}^{(l+1)} s_{ij}^{(l+1)}}$
If $\max_i \ v_i^{(l+1)} - v_i^{(l)}\ < \epsilon$, then stop; else $l = l + 1$ and go to step 1.

متن کامل مقاله

دریافت فوری ←

ISIArticles

مرجع مقالات تخصصی ایران

- ✓ امکان دانلود نسخه تمام متن مقالات انگلیسی
- ✓ امکان دانلود نسخه ترجمه شده مقالات
- ✓ پذیرش سفارش ترجمه تخصصی
- ✓ امکان جستجو در آرشیو جامعی از صدها موضوع و هزاران مقاله
- ✓ امکان دانلود رایگان ۲ صفحه اول هر مقاله
- ✓ امکان پرداخت اینترنتی با کلیه کارت های عضو شتاب
- ✓ دانلود فوری مقاله پس از پرداخت آنلاین
- ✓ پشتیبانی کامل خرید با بهره مندی از سیستم هوشمند رهگیری سفارشات