Pattern Recognition 44 (2011) 1750–1760

Contents lists available at ScienceDirect

Pattern Recognition

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

Fuzzy *C*-means based clustering for linearly and nonlinearly separable data

Du-Ming Tsai *, Chung-Chan Lin

Department of Industrial Engineering & Management, Yuan-Ze University, 135 Yuan-Tung Road, Nei-Li, Tao-Yuan, Taiwan, ROC

ARTICLE INFO

Article history:
Received 21 June 2010
Received in revised form
18 January 2011
Accepted 9 February 2011
Available online 16 February 2011

Keywords:
Clustering
Fuzzy *C*-means
Kernel fuzzy *C*-means
Distance metric

ABSTRACT

In this paper we present a new distance metric that incorporates the distance variation in a cluster to regularize the distance between a data point and the cluster centroid. It is then applied to the conventional fuzzy *C*-means (FCM) clustering in data space and the kernel fuzzy *C*-means (KFCM) clustering in a high-dimensional feature space. Experiments on two-dimensional artificial data sets, real data sets from public data libraries and color image segmentation have shown that the proposed FCM and KFCM with the new distance metric generally have better performance on non-spherically distributed data with uneven density for linear and nonlinear separation.

1. Introduction

Clustering is an unsupervised learning method to partition a collection of multivariate data points into meaningful groups, where all members within a group represent similar characteristics and data points between different groups are dissimilar to each other. It has been an important technique for pattern recognition, image processing and data mining. It has also been applied successfully in many fields such as marketing that finds groups of customers with similar purchasing behaviors, biology that groups unknown plants/animals into species, and medical image processing that divides an image into a few meaningful regions for diagnosis.

The similarity criterion for distinguishing the difference between data points is generally measured by distance. Two data points belong to the same group if they are close to each other. They are evidently from different groups if the distance between them is distinctly large. The success of a clustering algorithm is highly affected by the data structure including the cluster shape, cluster density and linear/nonlinear separability. The fuzzy *C*-means (FCM) algorithm [1] is one of the most popular techniques used for clustering. The effectiveness of the clustering method relies on the distance measure. The conventional FCM method uses the Euclidean distance as the similarity criterion that measures the distance between each data point $\boldsymbol{x}_i$ and a cluster centroid $\boldsymbol{v}_c$, i.e. $||\boldsymbol{x}_i-\boldsymbol{v}_c||^2$, with a weight w_{ic} which is inversely proportional to the distance. The Euclidean squared-norm distance makes FCM only suitable for clustering hyperspherically distributed data groups. In order to improve the performance of the conventional FCM, Wu and Yang [2] replaced the Euclidean norm with a normalized distance function $1-\exp(-\beta\|\boldsymbol{x}_i-\boldsymbol{v}_c\|^2)$, where β is a positive constant. Zhang and Chen [3,4] used $1-K(\boldsymbol{x}_i, \boldsymbol{v}_c)$ as the distance measure, where $K(\boldsymbol{x}_i,\boldsymbol{v}_c)$ is a kernel function. The normalized distance function proposed by Wu and Yang is only a special case of Zhang and Chen's method when a Gaussian function is used as the kernel. In a multi-dimensional space, some data features could be more critical than others. A feature-weighted distance [5,6] was proposed to improve the performance of FCM. The distance measure is given by $\sum_k \alpha_k[\boldsymbol{x}_i-\boldsymbol{v}_c]_k^2$, where $[\boldsymbol{x}_i-\boldsymbol{v}_c]_k$ is the difference of the kth feature between $\boldsymbol{x}_i$ and $\boldsymbol{v}_c$, and α_k is the assigned feature weight.

The conventional FCM only works for linearly separable data points. Girolami [7] proposed a kernel-based FCM by mapping the data in the observation space to a higher dimensional feature space so that nonlinear separation of clusters can be achieved. It uses the radial basis function (RBF) kernel to implicitly define the mapping function from data space to feature space. For a RBF kernel, the kernel-based FCM can be interpreted as replacing the Euclidean metric in the FCM algorithm by a probability metric. The choice of the RBF kernel bandwidth remains an open question in the paper. All the bandwidth values for the test data sets in the experiments were empirically determined. For the test samples with known class labels, the best bandwidth can be surely determined by an exhaustive search. To apply the kernel FCM to real data sets where the true class labels are not known, the best bandwidth value cannot be determined by trial-and-error or any search process since the true recognition rate is unknown. This problem is common in all unsupervised learning methods.

* Corresponding author. Fax: +886 3 463 8907.
E-mail addresses: iedmtsai@saturn.yzu.edu.tw, s929501@mail.yzu.edu.tw, s968902@mail.yzu.edu.tw (D.-M. Tsai).

0031-3203/$ - see front matter © 2011 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2011.02.009

Kim et al. [8] evaluated the performance of four kernel-based clustering methods including kernel *K*-means, kernel FCM, kernel average linkage algorithm and kernel mountain algorithm. The RBF function was used as the kernel in all four kernel clustering algorithms. The results in their experiments indicated each kernel clustering algorithm outperforms its conventional counterpart. The choice of the RBF bandwidth value that derived the superiority over the conventional methods for each individual test data set was not addressed in their paper.

Ng et al. [9] presented a spectral clustering method to improve the clustering results in the original data space. It consists of mapping the original space into a compact feature space by means of eigenvector decomposition, followed by *K*-means clustering in the new feature space to obtain better clustering results. The dominant eigenvectors are extracted from an affinity matrix constructed with elements $\exp(-\|\boldsymbol{x}_i-\boldsymbol{x}_j\|^2/2\sigma^2)$ for sample pairs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. The parameter value σ must be searched in a wide range with a pre-determined resolution, and the one that gives the tightest (smallest distortion) clusters in the new feature space is chosen. For each possible σ value, the whole clustering procedure including the calculation of eigenvectors from the large affinity matrix and then the *K*-means clustering must be repeated once. It is thus computationally very expensive. Zelnik-Manor and Perona [10] studied the parameter selections in spectral clustering. The bandwidth parameter is adaptively given by $\sigma_i\cdot\sigma_j=d(\boldsymbol{x}_i, \boldsymbol{x}_{K,i})\ d(\boldsymbol{x}_j, \boldsymbol{x}_{K,j})$ for sample pairs $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, where $\boldsymbol{x}_{K,i}$ is the *K*th neighbor of the individual data point $\boldsymbol{x}_i$, and $d(\cdot,\cdot)$ is some distance measure. The value of neighboring *K* has to be determined empirically, and the parameter setting process is computationally intensive. Von Luxburg [11] also recommended similar parameter setting as a rule of thumb for the choice of the bandwidth value in spectral clustering. Fred and Jain [12] proposed an evidence accumulation clustering (EAC) method for combining various existing clustering algorithms and/or the same clustering algorithm with various parameter values to obtain a partition that is better than individual clustering algorithms. The evidence accumulation technique maps the clustering ensemble into a new similarity measure between patterns by accumulating pairwise patterns with a voting mechanism. It can be expected that the application of the EAC method can lead to even better partitions of complex data sets if more powerful clustering algorithms are used in the combination.

The currently existing fuzzy *C*-means clustering methods both in the observed data space and in the mapped feature space basically consider only the Euclidean distance between each data point and every cluster centroid. It describes only hyperspherical clusters in data space or in feature space. Furthermore, the density of data points in a cluster could be distinctly different from other clusters in a data set. The conventional metric evaluates only the distance between two individual data points. It ignores the global distance variation for all data points in a cluster.

In this study, we add the distance variation of each individual data group to regularize the distance between a data point and the cluster centroid. The new distance metric is then applied to both the conventional FCM and the kernel FCM. The proposed distance metric can be better applied to non-hyperspherically shaped data with uneven densities for linear separation (in the observed data space) and nonlinear separation (in the mapped feature space). For the proposed distance metric in the kernel FCM, we also introduce a simple bandwidth selection rule for the RBF kernel when kernel FCM is used for clustering of unlabeled real data. Two-dimensional artificial data sets are first used to evaluate the robustness of the proposed clustering methods on cluster shape, cluster density and linear/nonlinear separability. Real data sets from public data libraries are then used to evaluate the clustering results. Finally, the application of the conventional FCM and KFCM and the proposed clustering methods to color image segmentation is also given.

This paper is organized as follows: Section 2 describes four versions of the fuzzy *C*-means clustering methods: conventional FCM, the proposed distance metric for FCM, the kernel FCM (KFCM), and the proposed distance metric for KFCM. The selection rule of bandwidth value of a given data set is also discussed in this section. Section 3 presents the experimental results on 2D artificial data sets, real data sets from public databanks and color image segmentation. The paper is concluded in Section 4.

2. Fuzzy *C*-means based clustering

In this section, we present four clustering methods based on fuzzy *C*-means. The conventional FCM and its formulation are first described so that the implication of the remaining three clustering methods can be correspondingly presented. Let $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\ldots,\boldsymbol{x}_N\}$ be a collection of unlabeled data points, and $\boldsymbol{x}_i\in\Re^d$. The goal of clustering is to partition the data set $\boldsymbol{X}$ into *C* intrinsic groups.

2.1. Fuzzy C-means (FCM) clustering

FCM partitions the data set $\boldsymbol{X}$ into *C* clusters by minimizing the errors in terms of the weighted distance of each data point $\boldsymbol{x}_i$ to all centroids of the *C* clusters. That is,

$$\text{Min } J_{\text{FCM}}=\sum_{c=1}^{C}\sum_{i=1}^{N} w_{ic}^{p}\|\boldsymbol{x}_i-\boldsymbol{v}_c\|^2$$

s.t.

$$\sum_{c=1}^{C} w_{ic}=1,\quad i=1,2,\ldots,N$$

where p is the exponent.

By using the Lagrange multipliers, we can solve for the weight w_{ic}. The weight w_{ic} and the centroid $\boldsymbol{v}_c$ can be updated by the expectation–maximization (E–M) algorithm:

E-step:

$$w_{ic}=1/\sum_{j=1}^{C}\left(\frac{d_{ic}^2}{d_{ij}^2}\right)^{1/(p-1)}\quad \text{for } i=1,2,\ldots,N \text{ and } c=1,2,\ldots,C$$

where

$$d_{ic}^2=\|\boldsymbol{x}_i-\boldsymbol{v}_c\|^2$$

M-step:

$$\boldsymbol{v}_c=\frac{\sum_{j=1}^{N} w_{jc}^{p}\cdot\boldsymbol{x}_j}{\sum_{j=1}^{N} w_{jc}^{p}}\quad \text{for } c=1,2,\ldots,C$$

The E–M algorithm recursively proceeds until a convergence condition is satisfied.

2.2. New distance metric for FCM

The conventional FCM only takes into account the Euclidean distances between individual data points and centroids. It ignores the distance variation of the data points in the same cluster. It thus may degrade the performance of FCM for data points with uneven densities or non-hyperspherical shapes in individual clusters.

In order to improve the effectiveness of FCM, a new metric that takes the distance variation in each cluster as the regularization of the Euclidean distance is proposed in this paper. The new distance