Pattern Recognition 46 (2013) 2538–2547

Contents lists available at SciVerse ScienceDirect

Pattern Recognition

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

Rough clustering using generalized fuzzy clustering algorithm

Jim Z.C. Lai, Eric Y.T. Juan*, Franklin J.C. Lai

Department of Computer Science and Engineering, National Taiwan Ocean University, Keelung 202, Taiwan, ROC

ARTICLE INFO

Article history:
Received 13 April 2012
Received in revised form
19 December 2012
Accepted 2 February 2013
Available online 13 February 2013

Keywords:
Rough *k*-means clustering
Nearest-neighbor search
Knowledge discovery
Soft computing

ABSTRACT

In this paper, we present a rough *k*-means clustering algorithm based on minimizing the dissimilarity, which is defined in terms of the squared Euclidean distances between data points and their closest cluster centers. This approach is referred to as generalized rough fuzzy *k*-means (GRFKM) algorithm. The proposed method solves the divergence problem of available approaches, where the cluster centers may not be converged to their final positions, and reduces the number of user-defined parameters. The presented method is shown to be converged experimentally. Compared to available rough *k*-means clustering algorithms, the proposed method provides less computing time. Unlike available approaches, the convergence of the proposed method is independent of the used threshold value. Moreover, it yields better clustering results than RFKM for the handwritten digits data set, landsat satellite data set and synthetic data set, in terms of validity indices. Compared to MRKM and RFKM, GRFKM can reduce the value of Xie–Beni index using the handwritten digits data set, where a lower Xie–Beni index value implies the better clustering quality. The proposed method can be applied to handle real life situations needing reasoning with uncertainty.

1. Introduction

Data clustering is used frequently in a number of applications, such as vector quantization (VQ) [1–4], pattern recognition [5], knowledge discovery [6], speaker recognition [7], fault detection [8], document collection [9,10], and web/data mining [11]. The main objective of data clustering is to divide dissimilar objects into different clusters and group similar ones into the same cell. In the past years, soft computing methodology such as rough sets is used in the mining step of KDD process, which refers to knowledge discovery in databases [12].

Rough clustering divides a set of data points into several rough clusters. A rough cluster consists of the lower approximation and upper approximation. Objects in the lower approximation of a cluster belong to this cluster only; while objects in the upper approximation will also be members of upper approximations of other clusters. Therefore rough clustering allows overlapping partitions, which can deal with uncertainty. Lingras and West [13] developed a rough *k*-means clustering algorithm using the reduced interpretation of rough sets. Rough *k*-means clustering uses points as cluster representatives and the squared Euclidean distance as the dissimilarity measure between a vector and cluster representative. The rough *k*-means clustering algorithm is numerical instable and depends on initial parameter selection [14]. To reduce instability, Mitra used a genetic algorithm to optimize the selection of parameters [14]. However, Peters used another approach to develop a modified rough *k*-means algorithm to avoid the instability problem [15]. As shown in Ref. [15], the modified rough *k*-means clustering algorithm still faces the instability problem in some cases. That is, the modified rough *k*-means clustering may not be converged for some values of ζ, where ζ is a threshold which controls the upper and lower approximations of a cluster. In Ref. [16], a method which integrates fuzzy and rough clustering, is developed to handle overlapping of clusters.

Like *k*-means clustering, an iterative process is required for rough *k*-means to generate a set of cluster representatives [17]. For *k*-means clustering, it is found that most of the centers are converged to their final positions and the majority of data points have few candidates to be selected as their closest centers [18,19]. It is also expected that rough *k*-means clustering will have the same behavior, if the corresponding process is converged. This characteristic may be used to reduce the computing time of rough *k*-means clustering.

In this paper, we will present an algorithm to solve the instability problem of rough *k*-means clustering through minimizing an objective function. The proposed approach is different from available methods which are more or less heuristic. This difference makes the proposed method stable, although it still faces the local minimum problem. The proposed method is

* Corresponding author. Tel.: +886 2 24622192x6621; fax: +886 2 24623249.
E-mail addresses: yjuan@mail.ntou.edu.tw, yjuan@ntou.edu.tw (E.Y.T. Juan).

0031-3203/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.patcog.2013.02.003

referred to as generalized fuzzy k-means clustering algorithm (GRFKM). GRFKM consists of two steps: partition step and new cluster center generation step in each iteration. GRFKM performs the partition step and new cluster center generation step repeatedly until all cluster centers are converged. The iterative process of GRFKM is similar to that of fuzzy k-means clustering algorithm. Therefore the characteristics of convergence and local minimum for fuzzy k-means clustering algorithm are also preserved by GRFKM. This paper is organized as follows. Section 2 describes related works. Section 3 presents the algorithms developed in this paper. Experimental results are presented in Section 4 and conclusions are given in Section 5.

2. Related works

In this section, rough k-means, modified rough k-means, and rough fuzzy k-means algorithms [13,15,16] will be briefly discussed. Let the set $S=\{\mathbf{X}_i\}$ consist of N data points. Rough k-means clustering divides the set S into k rough clusters. A rough cluster, which is defined similar to a rough set, consists of the lower approximation and upper approximation. A rough set for a data set S composes of the lower and upper approximations of the data set. When the lower and upper approximations of S are not equal, then the set S is said to be roughly definable. If the lower and upper approximations of S are the same, the set S is definable. The advantage of using rough sets is that rough set theory does not require any prior information such as apriori probability.

Denote R_j as the jth cluster, which consists of the lower approximation RL_j and upper approximation RU_j. A data point $\mathbf{X}$ in the lower approximation of a cluster is also member of the upper approximation of the same cluster. That is, $\mathbf{X}\in RL_j$ implies $\mathbf{X}\in RU_j$. A data object that is not in the lower approximation of any cluster belongs to the upper approximations of at least two clusters.

2.1. Rough k-means algorithm

Let $\mathbf{C}_j$ be the cluster representative of R_j and $d(\mathbf{X},\mathbf{C}_j)$ be the squared Euclidean distance between the data point X and cluster representative $\mathbf{C}_j$. Suppose that $\mathbf{C}_m$ is the closest cluster representative of $\mathbf{X}$. Let $T(\mathbf{X})=\{R_t\colon [d(\mathbf{X},\mathbf{C}_t)-d(\mathbf{X},\mathbf{C}_m)]\le\zeta \text{ and } t\ne m\}$, where ζ is a threshold. If $T(\mathbf{X})$ is an empty set, then $\mathbf{X}$ is in RL_m only; otherwise $\mathbf{X}$ belongs to RU_t for all $R_t\in T(\mathbf{X})$. Denote the boundary region of R_j as RB_j, where $RB_j=RU_j\backslash RL_j$. After each partition step of rough k-means clustering algorithm, the new cluster representatives are updated as follows:

$$\mathbf{C}_j=\begin{cases}\left(w_l\sum_{\mathbf{X}\in RL_j}\mathbf{X}\right)/|RL_j|+w_b\left(\sum_{X\in RB_j}\mathbf{X}\right)/|RB_j|, & \text{if } RB_j\ne\varnothing\\ \left(\sum_{\mathbf{X}\in RL_j}\mathbf{X}\right)/|RL_j|, & \text{otherwise}\end{cases}\tag{1}$$

where w_l and w_b are two parameters which define the importance of the lower approximation and boundary region and $\varnothing$ is an empty set. In expression (1), $|RL_j|$ and $|RB_j|$ denote the numbers of data objects in RL_j and RB_j, respectively. Now, we would like to present the rough k-means (RKM) algorithm as follows:

Rough k-means algorithm

(1) Assign randomly each data point of the data set S to the lower approximation of a cluster and set $RU_j=RL_j$, where $j=1, 2,\ldots, k$.
(2) Update the cluster representatives $\mathbf{C}_j$ ($j=1, 2,\ldots, k$) using Eq. (1).
(3) For each cluster $\mathbf{C}_j$, determine its nearest data object $\mathbf{X}_{nj}$. Set $RU_j=RL_j=\{\mathbf{X}_{nj}\}$. Let $S_n=\{\mathbf{X}_{n1}, \mathbf{X}_{n2},\ldots, \mathbf{X}_{nk}\}$.
(4) For each remaining data object $\mathbf{X}\in(S\backslash S_n)$:
  (4.1) Find its nearest cluster representative $\mathbf{C}_m$ and update $RU_m=RU_m\cup\{\mathbf{X}\}$.
  (4.2) Determine the set $T(\mathbf{X})=\{R_t\colon [d(\mathbf{X},\mathbf{C}_t)-d(\mathbf{X},\mathbf{C}_m)]\le\zeta$ and $t\ne m\}$, where $\mathbf{C}_m$ is the nearest cluster center of $\mathbf{X}$. If $T(\mathbf{X})$ is not empty, set $RU_t=RU_t\cup\{\mathbf{X}\}$ for all $R_t\in T(\mathbf{X})$. Otherwise update $RL_m=RL_m\cup\{\mathbf{X}\}$.
(5) Update the clustering representatives $\mathbf{C}_j$ ($j=1, 2,\ldots, k$) using Eq. (1).
(6) Repeat step (3) to step (5) until all cluster representatives are converged.

2.2. Modified rough k-means algorithm

The difference between rough k-means and modified rough k-means algorithms is that their approaches of updating cluster centers and determining the set $T(\mathbf{X})$ are different. Modified rough k-means algorithm uses the following expression to update cluster centers

$$\mathbf{C}_j=\left[w_l\left(\sum_{\mathbf{X}\in RL_j}\mathbf{X}\right)/|RL_j|+w_u\left(\sum_{\mathbf{X}\in RU_j}\mathbf{X}\right)/|RU_j|\right]\ \text{with } w_l+w_u=1\tag{2}$$

where $|RL_j|$ and $|RU_j|$ are the numbers of data objects in the lower and upper approximations of cluster R_j. For the modified rough k-means algorithm, $T(\mathbf{X})$ is determined by the following expression:

$$T(\mathbf{X})=\{R_t : [d(\mathbf{X},\mathbf{C}_t)/d(\mathbf{X},\mathbf{C}_{m,\mathbf{X}})]\le\varepsilon \text{ and } t\ne m\}\tag{3}$$

where ε is a threshold and $\mathbf{C}_m$ is the closest cluster representative of $\mathbf{X}$. At this stage, we would like to present the modified rough k-means (MRKM) algorithm as follows:

Modified rough k-means algorithm

(1) Assign randomly each data point of the data set S to the lower approximation of a cluster. Update the upper approximations of clusters by setting $RU_j=RL_j$, where $j=1, 2,\ldots, k$.
(2) Update the cluster representatives $\mathbf{C}_j$ ($j=1, 2,\ldots, k$) using Eq. (2).
(3) For each cluster $\mathbf{C}_j$, determine its nearest data object $\mathbf{X}_{nj}$ and set $RU_j=RL_j=\{\mathbf{X}_{nj}\}$. Let $S_n=\{\mathbf{X}_{n1}, \mathbf{X}_{n2},\ldots, \mathbf{X}_{nk}\}$.
(4) For each remaining data object $\mathbf{X}\in(S\backslash S_n)$:
  (4.1) Find its nearest cluster representative $\mathbf{C}_m$ and update $RU_m=RU_m\cup\{\mathbf{X}\}$.
  (4.2) Determine the set $T(\mathbf{X})=\{R_t\colon [d(\mathbf{X}, \mathbf{C}_t)/d(\mathbf{X}, \mathbf{C}_m)]\le\varepsilon$ and $t\ne m\}$, where ε is a threshold and $\mathbf{C}_m$ is the nearest cluster center of $\mathbf{X}$. If $T(\mathbf{X})\ne\phi$, set $RU_t=RU_t\cup\{\mathbf{X}\}$ for all $R_t\in T(\mathbf{X})$. Otherwise update $RL_m=RL_m\cup\{\mathbf{X}\}$.
(5) Update the cluster representatives $\mathbf{C}_j$ ($j=1, 2,\ldots, k$) using Eq. (2).
(6) Repeat step (3) to step (5) until all cluster representatives are converged or the maximum number of iterations is reached.

2.3. Rough fuzzy k-means algorithm

The rough fuzzy k-means algorithm involves the integration of fuzzy and rough sets. This algorithm incorporates fuzzy membership $u_{i,j}$ which indicates the degree of belongingness of a data object $\mathbf{X}_i$ with respect to cluster R_j in rough k-means clustering. The fuzzy membership $u_{i,j}$ can be calculated using the following equation:

$$u_{i,j}=\left((d(\mathbf{X}_i,\mathbf{C}_j))^{1/(q-1)}\sum_{l=1}^{k}\left(\frac{1}{d(\mathbf{X}_i,\mathbf{C}_l)}\right)^{1/(q-1)}\right)^{-1},\quad i=1 \text{ to } N \text{ and } j=1 \text{ to } k\tag{4}$$

where $d(\mathbf{X}_i,\mathbf{C}_j)$ is the squared Euclidean distance between data object $\mathbf{X}_i$ and cluster representative $\mathbf{C}_j$. Note here that N is the number of data objects and q is the fuzzifier parameter. For the rough fuzzy k-