Engineering Applications of Artificial Intelligence 56 (2016) 121–130

Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

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

Picture fuzzy clustering for complex data

Pham Huy Thong, Le Hoang Son [*]

VNU University of Science, Vietnam National University, 334 Nguyen Trai, Thanh Xuan, Hanoi, Viet Nam

ARTICLE INFO

Article history:
Received 24 April 2016
Received in revised form
9 August 2016
Accepted 9 August 2016

Keywords:
Complex data
Distinct structured data
Fuzzy clustering
Mix data type
Picture fuzzy clustering

ABSTRACT

Fuzzy clustering is a useful segmentation tool which has been widely used in many applications in real life problems such as in pattern recognition, recommender systems, forecasting, etc. Fuzzy clustering algorithm on picture fuzzy set (FC-PFS) is an advanced fuzzy clustering algorithm constructed on the basis of picture fuzzy set with the appearance of three membership degrees namely the positive, the neutral and the refusal degrees combined within an entropy component in the objective function to handle the problem of incomplete modeling in fuzzy clustering. A disadvantage of FC-PFS is its capability to handle complex data which include mix data type (categorical and numerical data) and distinct structured data. In this paper, we propose a novel picture fuzzy clustering algorithm for complex data called PFCA-CD that deals with both mix data type and distinct data structures. The idea of this method is the modification of FC-PFS, using a new measurement for categorical attributes, multiple centers of one cluster and an evolutionary strategy – particle swarm optimization. Experiments indicate that the proposed algorithm results in better clustering quality than others through clustering validity indices.

1. Introduction

Fuzzy clustering is used for partitioning dataset into clusters where each element in the dataset can belong to all clusters with different membership values (Bezdek et al., 1984). Fuzzy clustering was firstly introduced by Bezdek et al. (1984) under the name “Fuzzy C-Means (FCM)”. This algorithm is based on the idea of K-Means clustering with membership values being attached to the objective function for partitioning all data elements in the dataset into appropriate groups (Chen et al., 2016). FCM is more flexible than K-Means algorithm, especially in overlapping and uncertainty dataset (Bezdek et al., 1984). Moreover, FCM has many applications in real life problems such as in pattern recognition, recommender systems, forecasting, etc. (Son et al., 2012a, 2012b, Son et al., 2013, 2014; Thong and Son, 2014; Son, 2014a, 2014b; Son and Thong, 2015; Thong and Son, 2015; Son, 2015b, 2015c, 2016; Son and Tuan, 2016; Son and Hai, 2016; Wijayanto et al., 2016; Tuan et al., 2016; Thong et al., 2016; Tuan et al., 2016).

However, FCM still has some limitations regarding clustering quality, hesitation, noises and outliers (Ferreira and de Carvalho, 2012; De Carvalho et al., 2013; Thong and Son, 2016b). There have been many researches proposed to overcome these limitations; one of them is innovating FCM on advanced fuzzy sets such as the type-2 fuzzy sets (Mendel and John, 2002), intuitionistics fuzzy sets (Atanassov, 1986) and picture fuzzy sets (Cuong, 2014). *Fuzzy clustering algorithm on PFS* (FC-PFS) (Son, 2015a; Thong and Son, 2016b) is an extension of FCM with the appearance of three membership degrees of picture fuzzy sets namely the positive, the neutral and the refusal degrees combined within an entropy component in the objective function to handle the problem of incomplete modeling in FCM (Yang et al., 2004). FC-PFS was shown to have better accuracy than other fuzzy clustering schemes in the equivalent articles (Son, 2015a; Thong and Son, 2016b).

Nonetheless, a remark regarding the working flow of the FC-PFS algorithm extracted from our experiments through various types of datasets is the inefficiency of processing complex data which include mix data types and distinct structure data. Mix data are known as categorical and numerical data, which can be effectively processed with equipped kernel functions only (Ferreira and de Carvalho, 2012). Distinct structure data contains non-sphere structured data such as data scatter in a linear line or a ring types, etc. that prevent clustering algorithms to partition data elements into exact clusters. Almost fuzzy clustering methods, including FC-PFS, find them hard to deal with complex data. There have been many researches on developing new fuzzy clustering algorithms that employed dissimilarity distances and kernel functions to cope with complex data in (Cominetti et al., 2010; Hwang, 1998; Ji et al., 2013, 2012). However, they solved either mix data types or distinct structure data but not all of them so this

[*] Corresponding author.
E-mail addresses: thongph@vnu.edu.vn (P.H. Thong), sonlh@vnu.edu.vn, chinhson2002@gmail.com (L.H. Son).

http://dx.doi.org/10.1016/j.engappai.2016.08.009
0952-1976/© 2016 Elsevier Ltd. All rights reserved.

leaves the motivation for this paper to work on.

In this paper, we propose a novel *picture fuzzy clustering algorithm for complex data* called PFCA-CD that deals with both mix data type and distinct data structures. The idea of this method is the modification of FC-PFS, using a measurement for categorical attributes, multiple centers of one cluster and an evolutionary strategy - particle swarm optimization. Experiments indicate that the proposed algorithm results in better clustering quality than others through clustering validity indices.

The rest of the paper is organized as follows. Section 2 describes the background with literature review and some particular fuzzy clustering methods for complex data. Section 3 presents our proposed method. Section 4 validates the method on the benchmark UCI datasets. Finally, conclusions and further works are covered in last section.

2. Background

In this section, we firstly give an overview of the relevant methods for clustering complex data in Section 2.1. Sections 2.2–2.3 review two typical methods of this approach.

2.1. Literature review

The related works for clustering complex data is divided into two groups: mixed type of data including categorical and numerical data and distinct structure of data (Fig. 1).

In the first group, there have been many researches about clustering for both categorical and numerical data. Hwang (1998) extended the k-means algorithm for clustering large datasets including categorical values. Yang et al. (2004) used fuzzy clustering algorithms to partition mixed feature variables by giving a modified dissimilarity measure for symbolic and fuzzy data. Ji et al. (2012, 2013) proposed fuzzy k-prototype clustering algorithms combining the mean and fuzzy centroid to represent the prototype of a cluster and employing a new measure based on co-occurrence of values to evaluate the dissimilarity between data objects and prototypes of clusters. Chen et al. (2016) presented a soft subspace clustering of categorical data by using a novel soft feature-selection scheme to make each categorical attribute be automatically assigned a weight that correlates with the smoothed dispersion of the categories in a cluster. A series of methods based on multiple dissimilarity matrices to handle with mix data was introduced by De Carvalho et al. (2013). The main ideas of these methods were to obtain a collaborative role of the different dissimilarity matrices to get a final consensus partition. Although these methods can partition mixed data efficiently, they find it difficult to solve with complex distinct structure of data.

In the second group, many researchers tried to partition complex structure of data which had intrinsic geometry of non-sphere and non-convex clusters. Cominetti et al. (2010) proposed a method called DifFuzzy combining ideas from FCM and diffusion on graph to handle the problem of clusters with a complex nonlinear geometric structure. This method is applicable to a larger class of clustering problems which do not require any prior information on the number of clusters. Ferreira and de Carvalho (2012) presented kernel fuzzy clustering methods based on local adaptive distances to partition complex data. The main idea of these methods were based on a local adaptive distance where dissimilarity measures were obtained as sums of the Euclidean distance between patterns and centroids computed individually for each variable by means of kernel functions. Dissimilarity measure is utilized to learn the weights of variables during the clustering process that improves performance of the algorithms. However, this method could deal with numerical data only.

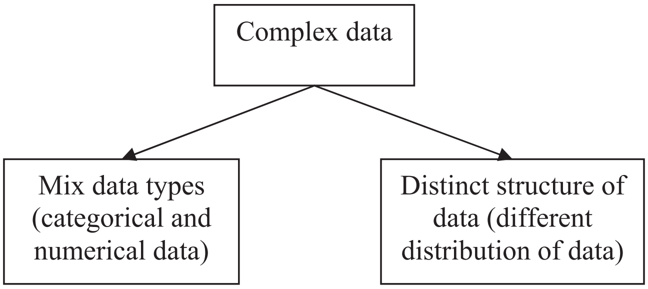
Complex data
Mix data types (categorical and numerical data)
Distinct structure of data (different distribution of data)

Fig. 1. Classification of methods dealing with complex data.

It has been shown that the DifFuzzy algorithm (Cominetti et al., 2010) and the fuzzy clustering algorithm based on multiple dissimilarity matrices (Dissimilarity) (De Carvalho et al., 2013) are two typical clustering methods in each group. Therefore, we will analyze these methods more detailed in the next sections.

2.2. DifFuzzy

DifFuzzy clustering algorithm (Cominetti et al., 2010) is based on FCM and the diffusion on graph to partition the dataset into clusters with a complex nonlinear geometric structure. Firstly, the auxiliary function is defined:

$$F(\sigma):\left(0,\,\infty\right)\to N \tag{1}$$

where $\sigma\in\left(0,\,\infty\right)$ be a positive number. The $i-th$ and $j-th$ nodes are connected by an edge if: $\|X_i-X_j\|<\sigma$. $F(\sigma)$ is equal to the number of components of the $\sigma-$ *neighborhood* graph which contain at least M vertices, where M is the mandatory parameter of DifFuzzy. $F(\sigma)$ begins from zero, and then increases to its maximum value, before settling back down to a value of 1.

$$C=\max_{\sigma\in\left(0,\infty\right)}F(\sigma), \tag{2}$$

$$\hat{\omega}_{i,j}(\beta)=\begin{cases}1\ if\ i\ and\ j\ are\ hard\ po\ int\ s\ in\ the\ same\ core\\ \quad clusters,\\ \exp\left(-\dfrac{\|X_i-X_j\|^2}{\beta}\right)otherwise,\end{cases} \tag{3}$$

where β is a positive real number. The function: L(β): $\left(0,\,\infty\right)\to\left(0,\,\infty\right)$ is:

$$L(\beta)=\sum_{i=1}^{N}\sum_{j=1}^{N}\hat{\omega}_{i,j}(\beta). \tag{4}$$

It has two well defined limits:

$$\lim_{\beta\to0}L(\beta)=N+\sum_{i=1}^{C}n_i\left(n_i-1\right)\ and\ \lim_{\beta\to\infty}L(\beta)=N^2, \tag{5}$$

where n_i corresponds to the number of points in the $i-th$ core cluster. DifFuzzy does this by finding β^* which satisfies the relation:

$$L\left(\beta^*\right)=\left(1-\gamma_i\right)\left(N+\sum_{i=1}^{C}n_i\left(n_i-1\right)\right)+\gamma_iN^2, \tag{6}$$

where $\gamma_1\in\left(0,\,1\right)$ is an internal parameter of the method. Its default value is 0.3. Then the auxiliary matrices are defined as follows.

$$W=\hat{W}\left(\beta^*\right). \tag{7}$$