ARTICLE IN PRESS
JID: KNOSYS [m5G;April 5, 2017;9:1]
Knowledge-Based Systems 000 (2017) 1–9
Contents lists available at ScienceDirect
Knowledge-Based Systems
journal homepage: www.elsevier.com/locate/knosys

Multiobjective fuzzy clustering approach based on tissue-like membrane systems

Hong Peng[a], Peng Shi[b], Jun Wang[c,*], Agustín Riscos-Núñez[d], Mario J. Pérez-Jiménez[d]

[a] School of Computer and Software Engineering, Xihua University, Chengdu, Sichuan, 610039, China
[b] School of Electrical and Electronic Engineering, University of Adelaide, Adelaide, SA 5005, Australia
[c] School of Electrical Engineering and Electronic Information, Xihua University, Chengdu, Sichuan, 610039, China
[d] Research Group of Natural Computing, Department of Computer Science and Artificial Intelligence, University of Seville, Sevilla, 41012, Spain

ARTICLE INFO

Article history:
Received 24 November 2016
Revised 21 February 2017
Accepted 29 March 2017
Available online xxx

Keywords:
Fuzzy clustering
Multiobjective clustering problem
Membrane computing
Tissue-like membrane system

ABSTRACT

Fuzzy clustering problem is usually posed as an optimization problem. However, the existing research has shown that clustering technique that optimizes a single cluster validity index may not provide satisfactory results on different kinds of data sets. This paper proposes a multiobjective clustering framework for fuzzy clustering, in which a tissue-like membrane system with a special cell structure is designed to integrate a non-dominated sorting technique and a modified differential evolution mechanism. Based on the multiobjective clustering framework, a fuzzy clustering approach is realized to optimize three cluster validity indices that can capture different characteristics. The proposed approach is evaluated on six artificial and ten real-life data sets and is compared with several multiobjective and singleobjective techniques. The comparison results demonstrate the effectiveness and advantage of the proposed approach on clustering the data sets with different characteristics.

1. Introduction

Clustering as a class of machine learning techniques has been widely used in many fields, such as pattern recognition, image processing, data mining and bioinformatics [1,2]. Clustering is the task of finding natural partitioning within a data set such that patterns within the same cluster are more similar than those within different clusters. Fuzzy c-means (FCM) [3] and k-means [4] are two of the most popular clustering algorithms, in which data clustering is regarded as an optimization problem and total cluster variance J_m is used as the objective function to be optimized. However, the two classical algorithms have a notable shortcoming: they easily fall into local minima and may not converge to the global minima [5]. To overcome this shortcoming, some global optimization techniques have been introduced to deal with data clustering problems in the past years, for example, simulated annealing (SA)-based [6], differential evolution (DE)-based [7–9], black hole-based [10], particle swarm optimization (PSO)-based [11–13], artificial bee colony (ABC)-based [14], genetic algorithms (GA)-based [15–17] and ant colony optimization (ACO)-based clustering algorithms [18] as well as a gravitational clustering algorithm [19]. In these clustering algorithms, a variety of cluster validity indices have been used to evaluate the goodness of partitioning obtained by them, such as Sym-index [20], *I*-index [21] and XB-index [22]. The existing works have indicated that these cluster validity indices have different characteristics, for example, XB-index is very suitable for processing the hyperspherical shaped clusters [8,9], while Sym-index that is more useful to detect symmetrical sharped clusters [23].

Most optimization-based clustering algorithms are single-objective because only a single validity measure is optimized. Note that a single validity measure can only reflect some intrinsic partitioning properties, for example, the compactness of clusters, the spatial separation between the clusters and the cluster's symmetry. Therefore, some clustering algorithms that use J_m as the objective function are only able to find compact hyperspherical and convex clusters like k-means. If clusters with different geometric shapes are present in the same data set, the clustering algorithms that use a single cluster validity index will fail to deal with the data set. Therefore, it is required to simultaneously optimize several cluster validity indices that can capture different data characteristics. Based on this consideration, data clustering should be viewed as a multiobjective optimization problem. Several works on multiobjective clustering have been published recently. Handl and Knowles [24] proposed a multiobjective clustering technique, *MOCK*, which can recognize the appropriate partitioning from the data sets that contain either hyperspherical shaped clusters or well-separated clusters. But, it fails to detect overlapping clusters

* Corresponding author.
E-mail address: wj.xhu@hotmail.com (J. Wang).

http://dx.doi.org/10.1016/j.knosys.2017.03.024

that have different shapes rather than hyperspheres. Another disadvantage of *MOCK* lies in its encoding, which can cause that the length of each chromosome will increase largely with the increase in the number of points. Faceli et al. [25] presented a clustering algorithm that combined cluster ensemble and multi-objective clustering techniques. Saha and Bandyopadhyay [26] presented a multiobjective clustering technique, called *VAMOSA*, where center-based encoding was used. Two cluster validity indices were optimized simultaneously: an Euclidean distance-based XB-index, and another point symmetry distance-based Sym-index. Experimental results indicated that *VAMOSA* can evolve the appropriate partitioning from data sets having clusters of any shape, size or convexity. The multiobjective clustering technique proposed in Saha and Bandyopadhyay [27], *GenClustMOO*, used a simulated annealing-based multiobjective clustering technique as the underlying optimization strategy. In Saha et al. [28], a multiobjective modified differential evolution-based fuzzy clustering, *MOmoDEFC*, has been addressed, in which both XB-index and FCM measure (J_m) were used as two objective functions. Simulation results showed that *MOmoDEFC* can optimize both the compactness and separation of clusters simultaneously. However, it may not be perfect to process the data set having cluster symmetry.

Membrane computing, initiated by Păun [29], is a class of distributed parallel computing models, known as membrane systems or P systems. The novel computing models were inspired from the structure and functioning of living cells as well as the cooperation of cells in tissues, organs and populations of cells [30–37]. In recent years, the inherent advantages and characteristics that membrane systems possess have attracted much attention on applications of membrane computing [38–43], for example, membrane algorithms for solving optimization problems. The research results on a variety of optimization problems have exhibited the potentiality of membrane computing models in the following three aspects: better convergence, stronger robustness and better balance between exploration and exploitation [44–49].

Based on the above consideration, the main motivation of this work is using membrane systems to develop a multiobjective optimization framework for fuzzy clustering problems. The role of the tissue-like membrane system stays in three aspects: (i) integrating differential evolution mechanism, (ii) realizing non-dominated sorting strategy, and (iii) realizing the coevolution between the objects in different cells. According to the multiobjective framework, three cluster validity indices, J_m, XB-index and Sym-index, are selected as objective functions, and a novel multiobjective fuzzy clustering approach is proposed in this paper, called *MOFC-TMS*. In Peng et al.[50], an evolution-communication membrane system has been used to propose a fuzzy cluster approach, called *Fuzzy-MC*. However, Fuzzy-MC is single-objective because only XB-index is considered as objective function to be optimized. In contrast to *Fuzzy-MC, MOFC-TMS* has two differences: (i) membrane systems are considered to solve multiobjective fuzzy clustering problems; (ii) a tissue-like membrane system with a special membrane structure is considered and a modification of differential evolution mechanism is designed according to the special structure. In addition, a single-objective approach that uses the special membrane structure and the modified differential evolution mechanism is implemented in simulation, which has better clustering performance over *Fuzzy-MC*. To the best of our knowledge, this is the first attempt to use a membrane computing model to solve multiobjective fuzzy clustering problems.

The rest of this paper is arranged as follows. Section 2 introduces multiobjective fuzzy clustering problems. Section 3 briefly reviews the definition and inherent mechanism of tissue-like membrane systems. In Section 4, a multiobjective clustering framework for fuzzy clustering is described in detail. In Section 5, experimental results carried out on some benchmark data sets are presented. Finally, conclusions are drawn in Section 6.

2. Problem statement

Data clustering in a d-dimensional Euclidean space is a process, which partitions n data points into several groups according to some similarity. Suppose that $X = \{X_1, X_2, \dots, X_n\}$ is a data set consisting of n unlabeled data points, where $X_i = (x_{i1}, x_{i2}, \dots, x_{id})$, $i = 1, 2, \dots, n$. A fuzzy clustering approach tries to find a fuzzy partitioning, $\{C_1, C_2, \dots, C_K\}$, such that the similarity of the data points in the same cluster is maximum and data points from different clusters differ as much as possible. Note that for a fuzzy partitioning, a data point can belong to all classes with a certain fuzzy membership degree for each class. Therefore, an appropriate partitioning matrix, $U = [u_{ij}]_{K\times n}$, needs to be evolved, where $u_{ij} \in [0, 1]$ denotes the membership grade of the jth element to the ith cluster. The fuzzy partitioning should maintain the following properties:

$$\begin{cases} 0 < \sum_{j=1}^{n} u_{ij} < n & \text{for } i = 1, 2, \dots, K \\ \sum_{i=1}^{K} u_{ij} = 1 & \text{for } j = 1, 2, \dots, n \\ \sum_{i=1}^{K}\sum_{j=1}^{n} u_{ij} = n \end{cases} \tag{1}$$

In the existing optimization-based clustering algorithms, fuzzy clustering problem is regarded as an optimization problem, however, most of them are single-objective because only a single cluster valid index is optimized. Note that a cluster valid index focuses mainly on some intrinsic partitioning property. However, a data set may have different geometric shapes, for example, the compactness of clusters, the spatial separation between the clusters and the cluster's symmetry. So, a single cluster valid index can fail to deal with data sets that have different geometric shapes. Therefore, fuzzy clustering problem should be posed as a multiobjective optimization problem, in which more objective functions (cluster valid indices) are optimized simultaneously. There are three cluster valid indices used widely in single-objective clustering algorithms: J_m, XB-index and Sym-index. The existing results have shown that the three indices can capture different data characteristics: (i) J_m can detect hyperspherical shaped clusters; (ii) XB-index can well detect compact and hyperspherical shaped clusters and emphasize the separation between two nearest clusters; (iii) Sym-index is more effective to detect symmetrical sharped clusters from the data set.

Since the three cluster valid indices can better capture the intrinsic characteristics of samples, they will be used as the objective functions to be optimized simultaneously in this work. Thus, fuzzy clustering problem can be formally defined as a multiobjective minimization problem

$$\begin{cases} \min\limits_{(z_1,\dots,z_K)} \; [f_1, f_2, f_3] \\ f_1(z_1, \dots, z_K) = J_m(z_1, \dots, z_K) \\ f_2(z_1, \dots, z_K) = XB(z_1, \dots, z_K) \\ f_3(z_1, \dots, z_K) = 1/Sym(z_1, \dots, z_K) \end{cases} \tag{2}$$

where $z_1, \dots, z_K$ are K parameters to be optimized, which denote K cluster centers of a partitioning.

In FCM, J_m is defined as follows

$$J_m(Z) = \sum_{i=1}^{K}\sum_{j=1}^{n} u_{i,j}^2 d^2(x_j, z_i) \tag{3}$$