Pattern Recognition Letters 68 (2015) 34–40

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

An automatic clustering algorithm inspired by membrane computing ☆

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

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

ARTICLE INFO

Article history:
Received 23 May 2014
Available online 24 August 2015

Keywords:
Membrane computing
Membrane systems
Tissue-like membrane systems
Automatic clustering
Membrane clustering algorithm

ABSTRACT

Membrane computing is a class of distributed parallel computing models. Inspired from the structure and inherent mechanism of membrane computing, a membrane clustering algorithm is proposed to deal with automatic clustering problem, in which a tissue-like membrane system with fully connected structure is designed as its computing framework. Moreover, based on its special structure and inherent mechanism, an improved velocity-position model is developed as evolution rules. Under the control of evolution-communication mechanism, the tissue-like membrane system cannot only find the most appropriate number of clusters but else determine a good clustering partitioning for a data set. Six benchmark data sets are used to evaluate the proposed membrane clustering algorithm. Experiment results show that the proposed algorithm is superior or competitive to three state-and-the-art automatic clustering algorithms recently reported in the literature.

1. Introduction

Data clustering as one of the most useful data mining techniques has been widely used in many fields, such as pattern recognition, image processing, web mining and biology [7,10,22,38]. Clustering will accomplish such a task that finds out the natural partition from a data set such that data points belonging to the same class are as similar to each other as possible whereas data points from two different classes share the maximum difference [13]. Partional clustering is a class of the most important clustering methods, which attempts to directly decompose the data set into several disjointed clusters according to some criteria [44]. The criteria commonly adopted in clustering is minimizing some measure of dissimilarity in the samples within each cluster and maximizing the disimilarity of different clusters. K-means is a widely used partitional clustering algorithm [16]. However, k-means has the following disadvantages: (1) it is sensitive to the initial cluster centers and easy to get stuck at the local optimal solutions; (2) it takes large time cost to find the global optimal solution when the number of data points is large; (3) it requires a priori specification of the number of clusters.

In recent years, a number of global optimization methods have been introduced to overcome the disadvantages of k-means, such as genetic algorithms (GA), simulated annealing (SA), ant colony optimization (ACO), particle swarm optimization (PSO) and differential evolution (DE) algorithm. The global search ability of the GA was first developed to find the optimal cluster centers for a data set [21]. The GA-based methods use two different coding schemes to express the clustering solutions: (i) using the chromosome directly to encode the cluster number that each data point belongs to [20]; (ii) using the chromosome to describe the cluster centers [2]. First scheme can suffer from huge searching space and high computing cost when the number of data points is very large. Thus, second scheme is commonly adopted by most of GA-based methods [3,19]. Although many GAs have shown good performance for finding the promising regions of the search space, most of them often have two drawbacks: premature convergence and lack of good local search ability. Thus, in order to overcome the problems above, other global searching techniques have been successively developed for data clustering problem. A PSO-based clustering method has been proposed in [17], where the PSO is used to find the optimal cluster centers. In [39], ACO has been introduced to process data clustering problem. Moreover, ACO and SA has been combined to solve clustering problem in [25], and a hybrid evolutionary algorithm based on PSO and ACO to find the optimal cluster centers has been also presented in [24]. In addition, a hybrid clustering method, which combines GA and EM (expectation maximization) to automatically determine the optimal cluster centers has been proposed in [23].

The clustering methods described above use the different global searching techniques to find the optimal cluster centers for a data set to be clustered. However, these clustering methods have a limit in practical application: the number of clusters needs to be

☆ This paper has been recommend for acceptance by M. A. Girolami.

* Corresponding author. Tel.: +8613683440486.
E-mail address: ph.xhu@hotmail.com, ph66@tom.com (H. Peng).

http://dx.doi.org/10.1016/j.patrec.2015.08.008
0167-8655/

determined a priori. In fact, it is difficult to specify the number of clusters in advance for most application. Thus, it becomes a challenge in such a situation in order to determine an appropriate number of clusters and provide a good partitioning for a data set automatically, that is, automatic clustering problem. In recent years, GA, PSO and DE have been used to deal with the automatic clustering problem. A clustering method that uses the GA to automatically evolve the clusters has been presented in [1]. This method uses a variable-length chromosome to express both the cluster centers and the number of clusters, and then achieves automatic clustering by evolving the two parts at the same time. In [27], a PSO-based automatic clustering method has been reported, which first uses the PSO to find the optimal number of clusters and then determines the corresponding cluster centers by using k-means algorithm. In addition, a variable-length GA to solve automatic fuzzy clustering problem has been developed in [37], while an automatic clustering algorithm based on an improved differential evolution has been presented in [6].

Membrane computing, as a class of distributed parallel computing models, is inspired from the structure and functioning of living cells as well as the cooperation of cells in tissues, organs and populations of cells [35,36]. The models are commonly called membrane systems or P systems. Over the past years, a variety of variants of membrane systems have been proposed [12,15,32–34,40,42,43], including membrane algorithms of solving the global optimization problems. In recent years, membrane algorithms have attracted much attention on applications of membrane computing [26]. The research results on a lot of global optimization problems have shown that compared to the existing evolutionary algorithms, membrane algorithms offer a more competitive method due to three advantages: better convergence, stronger robustness and better balance between exploration and exploitation [14,29–31,45]. Based on the above consideration, this paper proposes an automatic clustering algorithm that uses a tissue-like membrane system with fully connected structure to determine the most appropriate number of clusters and find a good partition for a data set to be clustered. Moreover, a modification of velocity-position model is developed according to its special structure and evolution-communication mechanism, which can accelerate the object evolution and enhance the diversity of objects in the system.

The rest of this paper is organized as follows. Section 2 states the problem to be solved and then presents a brief of introduction of tissue-like membrane systems. The proposed membrane clustering algorithm is described in detail in Section 3. Experimental results and analysis are provided in Section 4. Finally, Section 5 draws the conclusions.

2. Preliminaries

2.1. Data clustering problem

Data clustering in a D-dimensional Euclidean space is such a process, which partitions a data set consisted of n data points into K groups (clusters) according to some similarity measure. It is well-known that minimizing some similarity measure to find the natural partitioning on a non-uniform data set is a NP-hard problem essentially [11,22,24].

Assume that $X = \{X_1, X_2, \ldots, X_n\}$ is a data set of n unlabeled data points, where $X_i = (x_{i1}, x_{i2}, \ldots, x_{iD})$ is its ith data point. For the data set X, a partitional clustering algorithm tries to find a partitioning, $\{C_1, C_2, \ldots, C_K\}$, such that the similarity of the data points in the same cluster is maximum and data points from different clusters differ as far as possible.

K-means algorithm is a widely used clustering technique, which attempts to find the optimal cluster centers for determining a good partitioning of a data set. In order to determine the optimal cluster centers, therefore, a data clustering problem can be viewed as an optimization (minimization) problem. The objective function used in k-means is the following total mean square error:

$$J_m(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{X_j \in C_i} ||X_j - z_i||^2 \tag{1}$$

where $z_1, z_2, \ldots, z_K$ are the cluster centers of the partitioning, $C_1, C_2, \ldots, C_K$, respectively. Note that in k-means each cluster center is the average of samples in the corresponding cluster. However, in most evolutionary clustering algorithms and the proposed algorithm, the cluster center is a representative point of the corresponding cluster and it is often different from the average of samples. Moreover, the cluster centers, as the solutions of an optimization problem, are determined by the evolutionary clustering algorithms.

In recent years several clustering validity indexes have been proposed to evaluate the goodness of partitioning obtained by a clustering algorithm, such as *DI* index [9], *DB* index [8], *PBM* index [28] and *CS* measure [4]. The existing works have shown that J_m index can well capture only hyperspherical shaped clusters. However, data sets may have different shapes, spatial separations, densities and sizes. Compared with other clustering validity indexes, advantage of *CS* measure lies in the effectiveness of dealing with the clusters with different densities and sizes [4,5,7]. The *CS* measure is defined as follows:

$$CS(K) = \frac{\sum_{i=1}^{K} \left[\frac{1}{N_i} \sum_{X_l \in C_i} \max_{X_j \in C_i} ||X_l - X_j|| \right]}{\sum_{i=1}^{K} \left[\min_{1 \le j \le K, j \ne i} ||m_i - m_j|| \right]} \tag{2}$$

where m_i denotes average (vector) of samples in ith cluster and is calculated as follows:

$$m_i = \frac{1}{N_i} \sum_{X_j \in C_i} X_j \tag{3}$$

Generally, the lower *CS* measure under the constrain that $CS > 0$ means that the obtained partition is better, namely, the considered clustering algorithm gains a good clustering performance, and vice versa.

2.2. Tissue-like membrane systems

Tissue-like membrane systems are a kind of variants of membrane systems, which are inspired from the behavior of multiple single-membrane cells evolved in a common environment. A tissue-like membrane system can be logically viewed as a net, in which each cell is regarded as a processor that deals with the objects and communicates them between the cells along the channels assigned in advance. The object processing is completed by evolution rules while object communication is achieved by communication rules. We briefly review the definition and inherent mechanism of tissue-like membrane systems. More detailed descriptions of tissue-like membrane systems can be found in [12,35].

A tissue-like membrane system of degree q is a construct

$$\Pi = (O, w_1, \ldots, w_q, R_1, \ldots, R_q, R', i_0) \tag{4}$$

where

(1) O is a finite non-empty alphabet (of objects);
(2) $w_i (1 \le i \le q)$ is finite set of strings over O, which represents multiset of objects initially present in cell i;
(3) $R_i (1 \le i \le q)$ is finite set of evolution rules in cell i;
(4) R' is finite set of communication rules of the form $(i, u/v, j)$, which represents communication rule between cell i and cell j, $i \ne j, i, j = 1, 2, \ldots, q, u, v \in O^*$;
(5) i_0 indicates the output region of the system.