Membrane Clustering: A Novel Clustering Algorithm under Membrane Computing

Hong Peng[1], Jiarong Zhang[1], Jun Wang[2], Tao Wang[3],
Mario J. Pérez-Jiménez[4], Agustín Riscos-Núñez[4]

[1] Center for Radio Administration and Technology Development,
Xihua University, Chengdu 610039, China
ph.xhu@hotmail.com
[2] School of Electrical and Information Engineering,
Xihua University, Chengdu 610039, China
[3] School of Electrical Engineering,
Southwest Jiaotong University, Chengdu 610031, China
[4] Research Group of Natural Computing,
Department of Computer Science and Artificial Intelligence,
University of Seville, Sevilla 41012, Spain

Summary. Membrane computing (known as P systems) is a class of distributed parallel computing models, this paper presents a novel algorithm under membrane computing for solving the data clustering problem, called as membrane clustering algorithm. The clustering algorithm is based on a tissue-like P system with a loop structure of cells. The objects of the cells express the candidate cluster centers and are evolved by the evolution rules. Based on the loop membrane structure, the communication rules realize a local neighborhood topology, which helps the co-evolution of the objects and improves the diversity of objects in the system. The tissue-like P system can effectively search for the optimal clustering partition with the help of its parallel computing advantage. The proposed clustering algorithm is evaluated on four artificial data sets and six real-life data sets. Experimental results show that the proposed clustering algorithm is superior or competitive to classical k-means algorithm and several evolutionary clustering algorithms recently reported in the literature.

1 Introduction

Data clustering is a fundamental conceptual problem in data mining, which describes the process of grouping data into classes or clusters such that the data in each cluster share a high degree of similarity while being very dissimilar to data from other clusters [1]. Over the past years, a large number of clustering algorithms have been proposed [2, 3, 4], which can be divided roughly as two categories: hierarchical and partitional. Hierarchical clustering proceeds successively by either merging smaller clusters into larger ones or by splitting larger clusters. Partitional

clustering attempts to directly decompose a data set into several disjointed clusters based on similarity measure, for example, mean square error (MSE). Clustering algorithms have been used in a wide variety of areas, such as pattern recognition, machine learning, image processing, web mining [5, 6]. In the present study, the classical k-means algorithm [7, 8] has received wide attention because of the following two reasons: (i) k-means has been recently elected and listed among the top most influential data mining algorithms [9]; (ii) it is at the same time very simple and quite scalable, as it has linear asymptotic running time with respect to any variable of the problem. However, k-means is sensitive to the initial cluster centers and easy to get stuck at the local optimal solutions. Moreover, k-means takes large time cost to find the global optimal solution when the number of data points is large.

In recent years, some evolutionary algorithms have been introduced to overcome the shortcomings of k-means algorithm because of their global optimization capability. Several genetic algorithms (GA)-based clustering algorithms were proposed, which used the two different methods to express the clustering solutions respectively. The first method uses the chromosome directly to encode the cluster number that each data point belongs to [27, 28]. However, this method does not reduce the size of the search space and searching cost of the optimal solution when the data points proliferate. Another method uses a relatively indirect representation, in which the chromosome encodes the cluster centers and each data is subsequently assigned to the closest cluster center [10, 11, 12, 13, 14]. However, most of GA-based clustering algorithms can suffer from the degeneracy when numerous chromosomes represent the same solution. The degeneracy can lead to inefficient coverage of the search space as the same configurations of clusters are repeatedly explored. It is for this reason that some researchers developed the particle swarm optimization (PSO)-based or ant colony optimization (ACO)-based clustering algorithms. Kao et al. have proposed a hybrid technique based on combining the k-means and PSO for cluster analysis [15]. Shelokar et al. have introduced an evolutionary algorithm based on ACO for clustering problem [16]. Niknam et al. have presented a hybrid evolutionary optimization algorithm based on the combination of PSO and ACO for solving the clustering problem [17].

Membrane computing initiated by Păun [18] in 1998, is inspired by the structure and functioning of living cells as well as the interaction of living cells in tissues, organs or neural nets. Membrane computing is a novel class of distributed parallel computing models, and also known as P systems. The computing models usually have three key elements: membrane structure, multisets of objects and rules [19]. Generally, the multisets of objects are placed in compartments surrounded by membranes and evolved by some given rules. In recent years, a large number of variants have been proposed [20, 21, 22, 23]. These efforts have addressed the parallel computing advantage of P systems as well as the high effectiveness of solving a variety of difficult problems, especially, P systems can solve a number of NP-hard problems in linear or polynomial time complexity [24]. Moreover, membrane algorithms, as a variant of P systems, have demonstrated a powerful global

optimization performance [25, 26]. This paper focuses on application of membrane computing to data clustering. Our motivation is applying the specially designed elements and inherent mechanisms of P systems to achieve a novel clustering algorithm, called membrane clustering algorithm in this paper.

The rest of this paper is organized as follows. Section 2 gives a brief outline of tissue-like P systems. The proposed membrane clustering algorithm is presented in Section 3, and experimental results and analysis are provided in Section 4. Finally, Section 5 draws the conclusions.

2 Tissue-like P systems

The P systems first proposed are cell-like P systems in which the membranes are arranged as a rooted tree [18, 19], where the root expresses the skin of the cells (the outermost membrane) and the leafs represent elementary membranes (which do not contain any other membrane). Its biological inspiration is from the morphology of the cells, where small vesicles are surrounded by the large vesicles. For tissue-like P systems, tree-like structure is changed as a general graph. It is from the two biological inspirations: intercellular communication and collaboration between neurons. The intercellular communication is based on symport/antiport rules, which are introduced as the communication rules of tissue-like P systems. In symport rules, objects cooperate to traverse a membrane together in the same direction, whereas in the case of antiport rules, objects residing at both sides of the membrane cross it simultaneously but in opposite directions.

Formally, a tissue-like P system (of degree $q > 0$) with symport/antiport rules is a construct

$$\Pi = (O, w_1, \ldots, w_q, R_1, \ldots, R_q, R', i_0) \tag{1}$$

where

(1) O is a finite alphabet, whose symbols are called objects;
(2) $w_i (1 \leq i \leq q)$ is finite set of strings over O, which represents multiset of objects initially present in cell i;
(3) $R_i (1 \leq i \leq q)$ 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, $i \neq j$, $i, j = 1, 2, \ldots, q$, $u, v \in O^*$;
(5) i_0 indicates the output region of the system.

From membrane structure, a tissue-like P system can be viewed as a net implicitly, which consists of the q cells labeled by $1, 2, \ldots, q$ respectively. Here, each cell is an elementary membrane. Usually, the environment is labeled by 0. The communication rule of the form $(i, u/v, j)$ indirectly indicates synaptic connection between cell i and cell j. The communication rules determine a virtual graph, where the nodes are the cells and the edges indicate if it is possible for pairs of cells to communicate directly. The net structure provides the flexibility of expressing the needed structures from simple to complex when we deal with real-world problems.

In tissue-like P systems, multisets of objects of the q cells are described by $w_1, w_1, \ldots, w_q$. Suppose any multiset of objects over O is available in the environment.

Generally speaking, a tissue-like P system includes the rules of two types: evolution rules and communication rules. Each cell usually contains one or more evolution rules, while a communication rule is built between two different cells. In above definition, $R_i(1 \leq i \leq q)$ is finite set of evolution rules in cell i, whose rule is of the form $u \rightarrow v, u, v \in O^*$. The application of the rule means that u will be evolved to v. In most of the existing tissue-like P systems and variants, evolution rule of the form is based on string of objects. However, when we apply it to solve real-world problem, we should design the corresponding evolution rules according to domain knowledge of the real-world problem. The communication rule of the form $(i, u/v, j)$ is called as antiport rule. The communication rule $(i, u/v, j)$ can be applied over two cells labeled by i and j when u is contained in cell i and v is contained in cell j. The application of this rule means that the objects of the multisets represented by u and v are interchanged between the two cells. Note that if either $i = 0$ or $j = 0$ then the objects are interchanged between a cell and the environment. If one of u or v in above rule is empty, the rule is called as symport rule, for example, $(i, u/\lambda, j)$. The application of the rule means that u will be communicated form cell i to cell j.

In tissue-like P systems, as usual in the framework of membrane computing, every cell as a computing unit works in a maximally parallel way (a universal clock is considered here). In a computing step, each object in a cell can only be used for one rule (non-deterministically chosen when there are several possibilities), but any object which can participate in a rule of any form must do it, i.e, in each step we can apply a maximal set of rules.

A computation in a tissue-like P system of degree d is a sequence of steps which start with the cells $1, \ldots, q$ containing the multisets $w_1, \ldots, w_q$ and where, in each step, one or more rules are applied to the current multisets of symbol objects. A computation is successful if and only if it halts. When it halts, it produces a final result in output cell.

3 The proposed membrane clustering algorithm

In this section, we will present in detail the developed membrane clustering algorithm, a novel clustering algorithm under the framework of membrane computing, which is based on a tissue-like P system with a loop structure of cells. As usual, the designed tissue-like P system consists of several cells, each of which contains a object or multiple objects. The cells have some evolution rules to evolve the objects of the system, while communication rules between cell membranes are used to exchange and share the objects. Moreover, the loop structure of cells is indicated indirectly by the communication rules. The tissue-like P system can realize the co-evolution of objects among the cells under the control of evolution rules

and communication rules. The role of the tissue-like P system is to search for the optimal cluster centers for a data set to be clustered.

In the following, we first describe several basic components, and then provide the proposed tissue-like P system and membrane clustering algorithm.

3.1 Clustering measure

Suppose that data set D has n sample points, $x_1, x_2, \ldots, x_n$, $x_i \in R^d (i = 1, 2, \ldots, n)$, and is partitioned into k clusters, $C_1, C_2, \ldots, C_k$. Denote by $z_1, z_2, \ldots, z_k$ the corresponding cluster centers. If the distances of sample point x_i to cluster centers $z_p (p = 1, 2, \ldots, k)$ satisfy

$$||x_i - z_j|| \leq ||x_i - z_p||, p = 1, 2, \ldots, k \text{ and } j \neq p, \tag{2}$$

then sample point x_i is assigned to cluster C_j, $i = 1, 2, \ldots, n$.

Usually, partitional clustering algorithm searches for the optimal cluster centers in the solution space according to some clustering measure in order to solve data clustering problem. A commonly used clustering measure is

$$M(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x_j \in C_i} ||x_j - z_i||. \tag{3}$$

Generally, the smaller the M value, the higher the clustering quality. In this work, the clustering measure is also used to evaluate the objects of the system during object evolution. If the M value of an object is the smaller, the object is the better, otherwise, it is worse.

3.2 Membrane structure

The membrane clustering algorithm proposed in this paper is based on a tissue-like P system of degree q, which consists of q cells, shown in Fig. 1. The cells are labeled by $1, 2, \ldots, q$, respectively. The region labeled by 0 is the environment. In this work, the environment is also output region of the system. The directed lines in Fig. 1 indicate the communication of objects between the q cells. Moreover, the q cells will be arranged as a loop topology based on the communication rules described below. As usual in P system, the q cells, as parallel computing units, will run independently. In addition, the environment always stores the best object found so far in the system. When the system halts, the object in the environment will be regarded as the output of whole system.

3.3 Objects

In the tissue-like P system, each cell contains several objects. The role of the designed tissue-like P system is to find the optimal cluster centers for a data set,

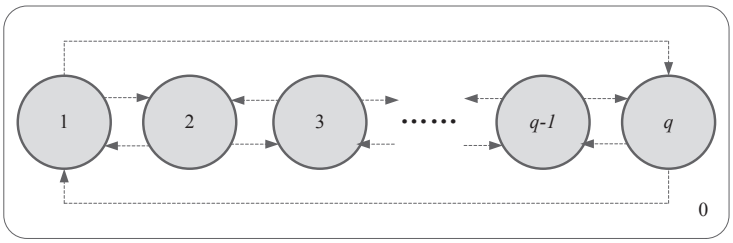

Fig. 1. Membrane structure of the designed tissue-like P system.

thus each object in cells will express a group of (candidate) cluster centers. Since data set D has k cluster centers and each cluster center is a d-dimensional vector, each object in the system is considered as a $(k \times d)$-dimensional real vector of the form

$$\mathbf{z} = (z_{11}, z_{12}, \ldots, z_{1d}, \ldots, z_{i1}, z_{i2}, \ldots, z_{id}, \ldots, z_{k1}, z_{k2}, \ldots, z_{kd})$$

where $z_{i1}, z_{i2}, \ldots, z_{id}$ are d components of ith cluster center z_i, $i = 1, 2, \ldots, k$. For simplicity, suppose that each cell has the same number of objects, which is denoted by m.

Initially, the system will randomly generates m initial objects for each cell. When an initial object $\mathbf{z}$ is generated, $(k \times d)$ random real numbers are produced repeatedly to form it with the constraint of

$$A_1 \leq z_{i1} \leq B_1, \ldots, A_j \leq z_{ij} \leq B_j, \ldots, A_d \leq z_{id} \leq B_d \tag{4}$$

where A_j and B_j are lower bound and upper bound of jth dimensional component of data points, respectively, $j = 1, 2, \ldots, d$.

3.4 Rules

The tissue-like P system includes the rules of two types: the evolution rules, which aim to evolve the objects in cells and the communication rules, which aim to exchange and share the objects. Evolution rules are used to evolve the objects associated with cluster centers, so the tissue-like P system is able to find the optimal cluster centers for a data set via the evolution of objects. Moreover, communication rules will realize the exchange and sharing of better objects between adjacent cells. Note that in each computing step, the communication rules are executed after the evolution rules. For each cell, the better objects communicated from its two adjacent cells form a subset of objects, called external pool, whose objects will participate its evolution of objects in next computing step (see Fig. 2). As usual in P systems, each cell as an independent computing unit runs in maximum parallel way under the control of a global clock.

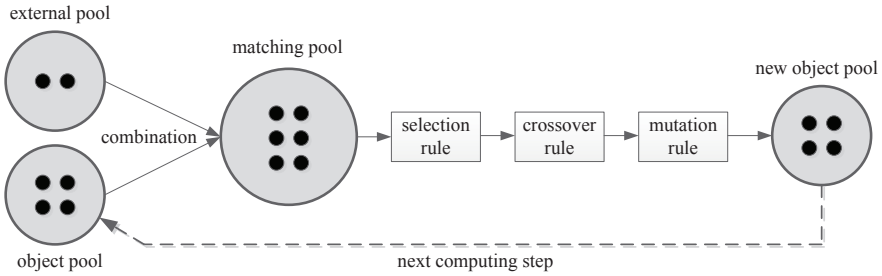

Fig. 2. The evolution procedure of objects in a cell.

Evolution rules

The role of evolution rules is to evolve the objects in cells to generate new objects used in next computing step. During the evolution, each cell maintains the same size (the number of objects). In this work, three known genetic operations (selection, crossover and mutation) [29, 30] are introduced as the evolution rules in cells. In a computing step, all objects (located in object pool) in each cell and the better objects (located in external pool) from its two adjacent cells constitute a matching pool. The objects in external pool are actually the better objects communicated from its two adjacent cells in previous computing step. The objects in matching pool will be evolved by executing selection, crossover and mutation operations in turn. In order to maintain the size of objects in each cell, truncation operation is used to constitute new object pool according to the M values of objects. The objects in new object pool will be regarded as the objects to be evolved in next computing step. Fig. 2 shows the evolution procedure of objects in a cell.

In this work, selection operation uses usual rotating wheel method, while crossover operation uses single-point crossover in which the position of crossover point is determined according to crossover probability p_c [31]. The single-point mutation is used to realize the mutations of objects. If v is a mutation point determined according to mutation probability p_m, its value becomes, after mutating,

$$v' = \begin{cases} v \pm 2 \times \delta \times v, & v \neq 0 \\ v \pm 2 \times \delta, & v = 0 \end{cases} \tag{5}$$

where the signs "+" or "-" occur with equal probability, and δ is a real number in the range [0,1], generated with uniform distribution.

Communication rules

The communication rules are used to exchange the objects between each cell and its two adjacent cells and update the best object found so far in the environment. The tissue-like P system designed in this paper involves the communication rules of two types:

(1) Antiport communication rule: $(i, \mathbf{z}/\mathbf{z}', j)$, $i, j = 1, 2, \ldots, q$. The rule indicates that object $\mathbf{z}$ is communicated from cell i to cell j and object $\mathbf{z}'$ is communicated from cell j to cell i.
(2) Symport communication rule: $(i, \mathbf{z}/\lambda, 0)$, $i = 1, 2, \ldots, q$. The rule expresses that object $\mathbf{z}$ is communicated from cell i to the environment.

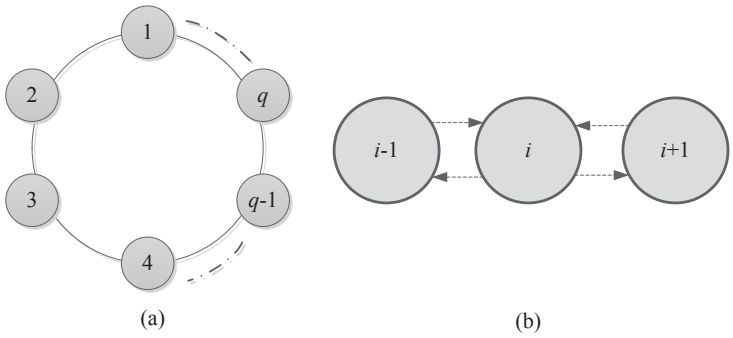

Fig. 3. A loop topology structure of cells and the communication relation between adjacent cells.

The communication rules impliedly indicate the connection relationship between cells. Fig. 3 shows the communication relation of objects between cells in the designed tissue-like P system. From a logical point of view, the communication relation shows that the cells form a loop topology, shown in Fig. 3(a). Meanwhile, this also reflects a neighborhood structure of the communication of objects, namely, each cell only exchanges and shares the objects with its two adjacent cells, shown in Fig. 3(b). After the objects are evolved, each cell (such as cell i) transmits its several best objects into adjacent cells (such as cells $i-1$ and $i+1$) and retrieves several best objects from adjacent cells (such as cells $i-1$ and $i+1$) by using the communication rule, constituting the matching pool of objects in next computing step. The special logical structure can bring the following benefits:

(1) The co-evolution of objects in the q cells can accelerate the convergence of the proposed clustering algorithm.
(2) The object sharing mechanism of the local neighborhood structure can enhance the diversity of objects in the entire system.

The communication of objects not only occurs between cells, but also appears between cell and the environment. The global best object found so far in whole system is stored always in the environment. After objects are evolved, each cell communicates its best object found in current computing step into the environment to update the global best object. The update strategy used in the tissue-like P

system is that if the communicated object is better than the global best object, the global best object is substituted, otherwise it is discarded.

3.5 Halt condition

In this paper, maximum execution step number is used as the halt condition of the tissue-like P system, that is, the tissue-like P system will continue to run until it reaches the maximum execution step number. When the system halts, the best object in the environment is regarded as the system output, which is the found optimal cluster centers.

3.6 The proposed clustering algorithm

According to the components discussed above, the designed tissue-like P system can be formally described as follows. It is a tissue-like P system of degree q,

$$\Pi = (Z_1, \ldots, Z_q, R_1, \ldots, R_q, R', i_o)$$

where

(1) Z_i is the set of m objects in cell i, where each object $\mathbf{z}$ is a $(k \times d)$-dimensional vector, $1 \leq i \leq q$;
(2) R_i is the finite set of evolution rules, $1 \leq i \leq q$. Each R_i contains three evolution rules: selection, crossover and mutation rules;
(3) R' is the finite set of communication rules with the following forms:
   (a) Antiport communication rule, $(i, \mathbf{z}/\mathbf{z}', j)$, $i, j = 1, 2, \ldots, q$, $i \neq j$. The rule is used to communicate the objects between an cell and its two adjacent cells;
   (b) Symport communication rule, $(i, \mathbf{z}/\lambda, 0)$, $i = 1, 2, \ldots, q$. The rule is used to communicate the objects between cell and the environment.
(4) $i_o = 0$ indicates that the environment is the output region of whole system.

Based on the tissue-like P system, the proposed membrane clustering algorithm is summarized in Table 1.

4 Experiment results and analysis

In this section, the proposed membrane clustering algorithm is evaluated on ten data sets and compared with classical k-means algorithm and several clustering algorithms based on evolutionary algorithms, including GA [10], PSO [15] and ACO [16]. In order to test the robustness of these clustering algorithms, we repeat the experiments 50 times for each data set.

Table 1. Membrane clustering algorithm: a clustering algorithm based on tissue-like P systems

Input parameters: Data set, D, the number of clusters, k, the number of cell, q, the number of objects in each cell, m, maximum execution step number, S_{max}, crossover rate, p_c, and mutation rate, p_m.
Output results: the optimal cluster centres, G.
Step 1. **Initialization**
for i=1 *to* q
for j=1 *to* m
Generate jth initial object for cell i, Z_{ij};
Partition all data points into clusters $C_1, C_2, \ldots, C_k$;
Compute the M value of the object, M_{ij};
end for
end for
Fill the global best object G using the best of all initial objects;
Set computing step $s = 0$;
Step 2. **Object evolution in cells**
for each cell i $(i = 1, 2, \ldots, q)$ in parallel *do*
Evolve all object Z_{ij} $(j = 1, 2, \ldots)$ in its mating pool using evolution rules;
Use truncation operation to maintain its m best objects;
for $j = 1$ *to* m
Partition all data points into clusters $C_1, C_2, \ldots, C_k$;
Compute the M value of the object, M_{ij};
end for
end for
Step 3. **Object communication between cells**
for each cell i $(i = 1, 2, \ldots, q)$ in parallel *do*
Transmit better objects in cell i to its two adjacent cells;
Receive better objects from its two adjacent cells into its mating pool;
Update G using the best object in cell i;
end for
Step 4. **Halt condition judgment**
if $s \leq S_{max}$ is satisfied
$s = s + 1$;
goto **Step 2**;
end if
The system exports the global best object G in the environment and halts;

4.1 Data sets

In the experiments, two kinds of data are used to evaluate these clustering algorithms. The first is the four manually-generated data sets used in the existing literatures, *AD_5_2*, *Data_9_2*, *Square_4* and *Sym_3_22*, shown in Fig. 4. The second is the six real-life data sets provided in UCI [32], including the *Iris*, *BreastCancer*, *Newthyroid*, *LungCancer*, *Wine* and *LiveDisorder*.

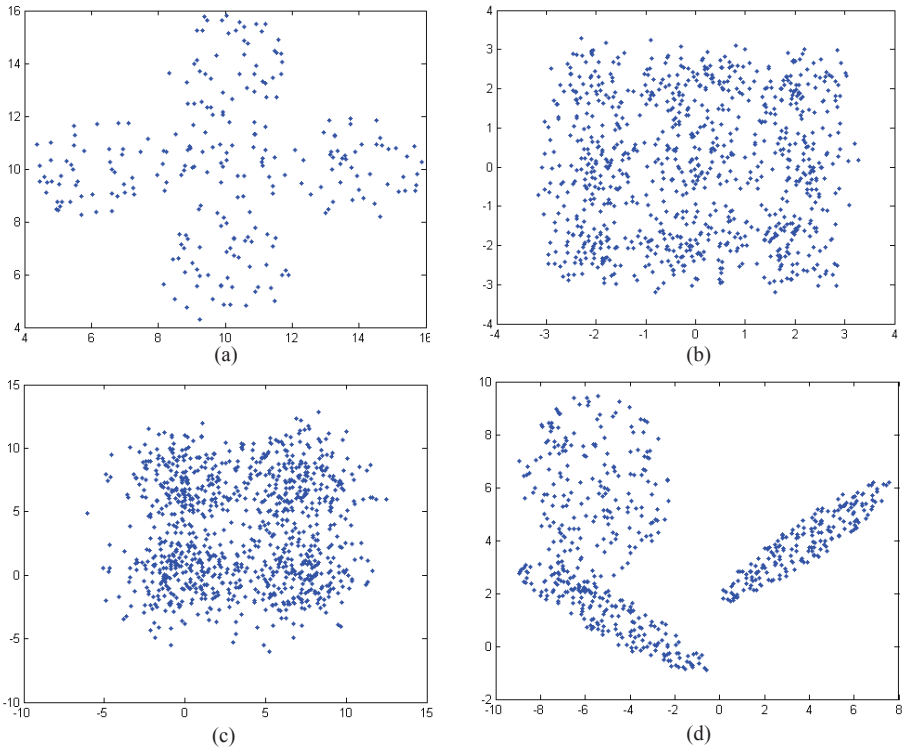

Fig. 4. Four artificial data sets: (a) *AD_5_2*; (b) *Data_9_2*; (c) *Square_4*; (d) *Sym_3_22*.

- *AD_5_2*. This data set consists of 250 two-dimensional data points distributed over five spherically shaped clusters. The clusters present in this data set are highly overlapping, each consisting of 50 data points. This data set is shown in Fig. 4(a).
- *Data_9_2*. This data set consists of 900 two-dimensional data points distributed over nine spherically shaped clusters. The clusters present in this data set are highly overlapping. This data set is shown in Fig. 4(b).
- *Square_4*. This data set consists of 1000 data points distributed over four squared clusters. This data set is shown in Fig. 4(c).

- *Sym_3_22*. This data set consists of 600 two-dimensional data points distributed over three clusters, where first and second clusters are spherically shaped while third cluster is elliptically shaped, each consisting of 200 data points. This data set is shown in Fig. 4(d).
- *Iris*. This data set consists of 150 data points distributed over three clusters. Each cluster consists of 50 points. This data set represents different categories of irises characterized by four feature values in centimeters: the sepal length, sepal width, petal length and the petal width [33]. This data set has three classes, namely, Setosa, Versicolor and Virginica, among which the last two classes have a large amount of overlap while the first class is linearly separable.
- *BreastCancer*. This data set consists of 683 sample points. Each pattern has nine features corresponding to clump thickness, cell size uniformity, cell shape uniformity, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. There are two categories in the data: malignant and benign. The two classes are known to be linearly separable.
- *Newthyroid*. The original database from where it has been collected is titled as thyroid gland data ("normal", "hypo" and "hyper" functioning). Five laboratory tests are used to predict whether a patient's thyroid belongs to the class euthyroidism, hypothyroidism or hyperthyroidism. There are a total of 215 instances and the number of attributes is five.
- *LungCancer*. The data consists of 32 instances having 56 features each. The data describes three types of pathological lung cancers.
- *Wine*: This is a wine recognition data consisting of 178 instances with 13 features resulting from a chemical analysis of wines grown in the same region in Italy but derived from three different cultivars. The analysis determined the quantities of 13 constituents found in each of the three types of wines.
- *LiveDisorder*. This data set contains 345 instances with six features each. The data has two categories. The first five variables are all blood tests, which are thought to be sensitive to liver disorders that might arise from excessive alcohol consumption.

4.2 Setup

In the experiments, the proposed membrane clustering algorithm will be compared with k-means and three evolutionary clustering algorithms recently reported in the literatures, including GA, PSO and ACO. These algorithms are implemented in Matlab 7.1 according to the following parameters:

- Tissue-like P systems. Each cell contains 100 objects and communicates its first five best objects into two adjacent cells. The maximum computing step number is chosen to be 200. In the implementation, evolution rules uses the adaptive crossover probability p_c and mutation probability p_m. In order to study performances of tissue-like P systems of different degrees, four cases are considered in the experiments: $q = 4, 8, 16, 20$.

- GA [10]. The rotating wheel method, single-point crossover and single-point mutation are used, where the crossover and mutation probabilities, p_c and p_m, are chosen to be 0.8 and 0.001 respectively. Let the population size be $N_{swarm} = 100$ and maximum iteration number be $t_{max} = 200$.
- PSO [15]. The ω uses a linear decreasing inertia weight, where $\omega_{min} = 0.4$ and $\omega_{max} = 0.9$. $c_1 = c_2 = 2.0$, the population size $NP = 100$, and maximum iteration number is 200.
- ACO [16]. The best parameter values are $\gamma_1 = \gamma_2 = 1.0$ and $\rho = 0.99$.

4.3 Experimental results

Table 2. The performance comparisons of tissue-like P systems of different degrees.

Data sets	4 cells	8 cells	16 cells	20 cells
AD_5_2	327.01	326.94	326.44	326.94
	±0.0944	±0.0277	±0.0105	±0.0312
Data_9_2	591.11	591.12	591.06	591.03
	±0.1331	±0.0510	±0.0280	±0.0537
Square_4	2380.25	2380.26	2379.74	2380.00
	±0.1334	±0.0956	±0.0189	±0.0729
Sym_3_22	1248.31	1248.11	1247.72	1248.05
	±0.3156	±0.0554	±0.0105	±0.0333
Iris	96.84	96.81	96.75	96.77
	±0.0751	±0.0435	±0.0428	±0.0361
BreastCancer	2974.24	2971.14	2970.24	2969.06
	±1.5431	±1.5287	±1.1225	±1.0970
Newthyroid	1885.69	1870.37	1869.29	1871.18
	±14.3773	±1.7355	±0.9215	±2.2496
LungCancer	124.69	124.69	124.69	124.69
	±0.0045	±0.0012	±0.0011	±0.0035
Wine	16309.01	16303.42	16292.25	16301.97
	±2.5053	±1.9595	±0.1529	±2.8563
LiveDisorder	9860.54	9859.02	9851.78	9857.08
	±5.7239	±0.5116	±0.0347	±0.1043

In the experiments, we realize four tissue-like P systems with degrees 4, 8, 16 and 20 respectively. The aim is to evaluate the effects of the number of cells (i.e., different degrees) on clustering quality. The four tissue-like P systems are applied to find out the optimal cluster centers for the ten data sets respectively. In this work, the M value is also used to measure the clustering quality of each clustering algorithm. Considering that the evolution rules in the designed tissue-like P system include stochastic mechanism, we independently execute the tissue-like P systems of the four degrees 50 times on each data set, and then compute their mean values and standard deviations of the 50 runs. The mean values are used to illustrate the

average performance of the algorithms while standard deviations indicate their robustness. Table 2 provides experimental results of the tissue-like P systems of four degrees on ten data sets respectively. The results of degrees 16 and 20 are better than those of other two degrees, namely, lower mean values and smaller standard deviations. It can be further observed that the tissue-like P system with degree 16 obtains the smallest mean values and standard deviations on most of data sets. The results illustrate that the tissue-like P system with degree 16 has good clustering quality and high robustness.

Table 3. The results obtained by the algorithms for 50 runs on the ten data sets.

Data sets	P systems	GA	PSO	ACO	k-means
AD_5_2	326.44	332.31	326.44	326.45	332.47
	±0.0105	±0.4792	±0.0128	±0.0344	±3.1286
Data_9_2	591.06	593.7251	591.14	591.42	623.57
	±0.0280	±0.2635	±0.0303	±0.0372	±3.1326
Square_4	2379.74	2380.33	2379.74	2379.79	2386.00
	±0.0189	±0.6319	±0.0226	±0.0428	±4.5217
Sym_3_22	1247.72	1249.36	1247.72	1247.75	1255.45
	±0.0105	±1.2163	±0.0149	±0.0315	±3.8725
Iris	96.75	99.83	97.23	97.25	104.11
	±0.0428	±5.5239	±0.3513	±0.4152	±12.4563
BreastCancer	2970.24	3249.26	3050.04	3046.06	3251.21
	±1.1225	±229.734	±110.801	±90.500	±251.143
Newthyroid	1869.29	1875.11	1872.51	1872.56	1886.25
	±0.9215	±13.5834	±11.0923	±11.1045	±16.2189
LungCancer	124.69	129.52	127.23	127.31	139.40
	±0.0011	±4.4961	±1.1528	±1.2936	±7.3136
Wine	16292.25	16298.42	16292.25	16292.25	16312.43
	±0.1529	±2.1523	±0.1531	±0.1672	±9.4269
LiveDisorder	9851.73	9856.14	9851.73	9851.74	9868.32
	±0.0347	±1.9523	±0.0356	±0.0692	±7.9274

In order to further evaluate clustering performance, the proposed membrane clustering algorithm is compared with GA-based, PSO-based and ACO-based clustering algorithms as well as classical k-means algorithm. Tables 3 gives the comparison results of the tissue-like P system of degree 16 with other four clustering algorithms on the ten data sets, respectively. The comparison results show that the tissue-like P system provides the optimum average value and smallest standard deviation in compare to those of other algorithms. For instance, the results obtained on the *AD_5_2* show that the tissue-like P system converges to the optimum of 326.4478 at almost times and PSO reaches to 326.44 in most of runs, while ACO, GA and k-means attain 326.45, 322.31 and 332.47 respectively. The standard deviations of M values for the tissue-like P system, PSO and ACO are 0.0105, 0.0128 and 0.0344 respectively, which significantly are smaller than other two algorithms.

For the results on the *Iris*, the optimum value is 96.75, which is obtained in most of runs of the tissue-like P system, however, other four algorithms fail to attain the value even once within 50 runs. The results on the *Newthyroid* also show that the tissue-like P system provides the optimum value of 1869.29 while the PSO, ACO, GA and k-means obtain 1872.51, 1872.56, 1875.11 and 1886.25 respectively. In addition, the tissue-like P system obtains smallest standard deviation on each data set in compare to other four algorithms, which illustrates that it has high robustness.

The Wilcoxon's rank sum test is a nonparametric statistical significance test for independent samples. The statistical significance test has been conducted at the 5% significance level in the experiments. We create five groups for the ten data set, which are corresponding to the five clustering algorithms (tissue-like P system, GA, PSO, ACO and k-means) respectively. Each group consists of the M values produced by 50 consecutive runs of the corresponding algorithms. In order to illustrate the goodness is statistically significant, we have completed a statistical significance test for these clustering algorithms. Table 4 gives the p-values provided by Wilcoxon's rank sum test for comparison of two groups (one group corresponding to the tissue-like P system and another group corresponding to some other method) at a time. The null hypothesis assumes that there is no significant difference between the mean values of two groups, whereas there is significant difference in the mean values of two groups for the alternative hypothesis. It is evident from Table 4 that all p-values are less than 0.05 (5% significance level). This is a strong evidence against the null hypothesis, establishing significant superiority of the proposed membrane clustering algorithm.

Table 4. The results of p-values produced by Wilcoxon's rank sum test.

Data sets	GA	PSO	ACO	k-means
AD_5_2	4.1321×10^{-3}	2.3256×10^{-2}	2.6351×10^{-2}	3.4273×10^{-3}
Data_9_2	4.0536×10^{-3}	2.2734×10^{-2}	2.7932×10^{-2}	3.2963×10^{-3}
Square_4	3.9275×10^{-3}	2.1482×10^{-2}	2.8175×10^{-2}	3.5387×10^{-3}
Sym_3_22	3.7894×10^{-3}	2.4357×10^{-2}	2.8529×10^{-2}	3.4416×10^{-3}
Iris	4.0968×10^{-3}	3.5823×10^{-2}	3.2634×10^{-2}	3.6528×10^{-3}
BreastCancer	3.9235×10^{-3}	2.9527×10^{-2}	2.8192×10^{-2}	3.4632×10^{-3}
Newthyroid	3.8864×10^{-3}	2.5162×10^{-2}	2.9355×10^{-2}	3.5381×10^{-3}
LungCancer	3.8575×10^{-3}	2.7346×10^{-2}	2.7358×10^{-2}	3.5138×10^{-3}
Wine	3.7639×10^{-3}	3.2189×10^{-2}	2.7963×10^{-2}	3.6348×10^{-3}
LiveDisorder	3.8398×10^{-3}	2.4671×10^{-2}	2.8846×10^{-2}	3.5822×10^{-3}

5 Conclusion

In this paper, we discuss a membrane clustering algorithm, a novel clustering algorithm under the framework of membrane computing. Distinguished from the

existing evolutionary clustering techniques, two inherent mechanisms of membrane computing are exploited to realize the membrane clustering algorithm, including evolution and communication mechanisms. For this purpose, a tissue-like P system consisting of q cells is designed, in which each cell as parallel computing unit runs in maximally parallel way and each object of the system expresses a group of candidate cluster centers. Moreover, the communication rules impliedly realize a local neighborhood structure, namely, each cell exchanges and shares the best objects with its two adjacent cells. Under the control of evolution and communication mechanisms of objects, the tissue-like P system is able to search for the optimal cluster centers for a data set to be clustered. In addition, the local neighborhood structure can guide the exploitation of the optimal object and enhance the diversity of evolution objects. Therefore, the membrane clustering presented in this paper can be viewed as a successful instance for building a bridge between membrane computing and data clustering.

Acknowledgements

This work was partially supported by the National Natural Science Foundation of China (Grant No. 61170030), the Chunhui Project Foundation of the Education Department of China (Nos. Z2012025 and Z2012031), and the Sichuan Key Technology Research and Development Program (No. 2013GZX0155), China.

References

1. J.A. Hartigan, *Clustering Algorithm*, New York: Wiley, 1975.
2. A.K. Jain, R.C. Dubes, *Algorithms for Clustering Data*, Prentice-Hall, Engiewood Cliffs, NJ, 1988.
3. R. Xu, D. Wunsch, Survey of clustering algorithm, IEEE Trans. Neural Networks 16(3) (2005) 645-678.
4. A.K. Jain, Data clustering: 50 years beyond k-means, Pattern Recognition Letters 31 (2010) 651-666.
5. B. Everitt, S. Landau, M. Leese, *Cluster Analysis*, Arnold, London, 2001.
6. S. Saha, S. Bandyopadhyay, A symmetry based multiobjective clustering technique for automatic evolution of clusters, Pattern Recognition 43 (2010) 738-751.
7. T. Kanungo, D. Mount, N.S. Netanyahu, C. Piatko, R. Silverman, A. Wu, An efficient k-means clustering algorithm: analysis and implementation, IEEE Trans. Pattern Anal. Mach.Intell. 24(7) (2002) 881-892.
8. D. Steinley, K-meams clustering: A half-century synchesis, British Journal of Mathematical and Statistics Psychology 59(34) (2006) 1-34.
9. X. Wu, Top ten algorithms in data mining, Taylor & Francis, 2009.
10. S. Bandyopdhyay, U. Maulik, An evolutionary technique based on k-means algorithm for optimal clustering in RN, Inf. Sci. 146 (2002) 221-237.
11. S. Bandyopdhyay, S. Saha, GAPS: a clustering method using a new point symmetry-based distance measure, Pattern Recognition 40 (2007) 3430-3451.

12. M. Laszlo, S. Mukherjee, A genetic algorithm that exchanges neighboring centers for k-means clustering, Pattern Recognition Lett. 28 (2007) 2359-2366.
13. D. Chang, X. Zhang, C. Zheng, A genetic algorithm with gene rearrangment for k-means clustering, Pattern Recognition 42 (2009) 1210-1222.
14. C.D. Nguyen, K.J. Cios, GAKREM: A novel hybrid clustering algorithm, Information Sciences 178 (2008) 4205-4227.
15. Y.T. Kao, E. Zahara, I.W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 34(3) (2008) 1754-1762.
16. P.S. Shelokar, V.K. Jayaraman, B.D. Kulkarni, An ant colony approach for clustering, Analytica Chimica Acta 509(2) (2004) 187-195.
17. T. Niknam, B. Amiri, An efficient hybrid approach based on PSO, ACO and k-means for cluster analysis, Applied Soft Computing 10 (2010) 183-197.
18. Gh. Păun, Computing with membranes, Journal of Computer System Sciences 61(1) (2000) 108-143.
19. Gh. Păun, G. Rozenberg, A. Salomaa, *The Oxford Handbook of Membrane Computing*, Oxford University Press, New York, 2010.
20. M. Ionescu, Gh. Păun, T. Yokomori, Spiking neural P systems, Fundameta Informaticae 71(2-3) (2006) 279-308.
21. R. Freund, Gh. Păun, M.J. Pérez-Jiménez, Tissue-like P systems with channel-states, Theoretical Computer Science 330(1) (2005) 101-116.
22. H. Peng, J. Wang, M.J. Pérez-Jiménez, H. Wang, J. Shao, T. Wang, Fuzzy reasoning spiking neural P system for fault diagnosis, Information Sciences 235 (2013) 106-116.
23. J. Wang, P. Shi, H. Peng, Mario J. Pérez-Jiménez, T. Wang, Weighted fuzzy spiking neural P systems, IEEE Transactions on Fuzzy Systems 21(2) (2013) 209-220.
24. Gh. Păun, M.J. Pérez-Jiménez, Membrane computing: Brief introduction, recent results and applications, BioSystems 85 (2006) 11-22.
25. L. Huang, I. Suh, A. Abraham, Dynamic mul-objective optimization based on membrane computing for control of time-varying unstable plants, Information Sciences, 181(11) (2011) 2370-2391.
26. G. Zhang, J. Cheng, M. Gheorghe, Q. Meng, A hybrid approach based on different evolution and tissue membrane systems for solving constrained manufacturing parameter optimization problems, Applied Soft Computing, 13(3) (2013) 1528-1542.
27. C.A. Murthy, N. Chowdhury, In search of optimal clusters using genetic algorithms, Parttern Recognition Letters 17 (1996) 825-832.
28. U. Maulik, S. Bandyopadhyay, Genetic algorithm based clustering technique, Pattern Recognition 33 (2000) 1455-1465.
29. E. Falkenauer, Genetic Algorithms and Grouping Problems, John Wiley & Sons, 1998.
30. L. Davis, Handbook of Genetic Algorithms, Van Nostrand Reinhold, 1991.
31. Z. Michalewicz, *Genetic Algorithm + Data Structure = Evolution Program*, Springer, New York, 1996.
32. http://www.ics.uci.edu/∼mlearn/MLRepository.html.
33. R.A. Fisher, The use of multiple measurements in taxonomic problems, Ann. Eugenics 3 (1936) 179-188.
34. S.K. Pal, D.D. Majumder, Fuzzy sets and decision making approaches in vowel and speaker recognition, IEEE Trans. Systems, Man Cybernet. SMC-7 (1977) 625-629.
35. M. Clerc, J. Kennedy, The particle swarm explosion stability and convergence in a multi-dimensional complex space, IEEE Trans. Evolutionary Comput. 6(1) (2002) 58-73.