A Novel Membrane Clustering Algorithm Based on Tissue-like P system

Huaning Yan, Laisheng Xiang, Xiyu Liu, Jie Xue
School of Management Science and Engineering
Shandong Normal University, Jinan, China
*Corresponding author, e-mail: sdxyliu@163.com

Abstract
Clustering is a process of partitioning data points into different clusters due to their similarity, as a powerful technique of data mining, clustering is widely used in many fields. Membrane computing is a computing model abstracting from the biological area, these computing systems are proved to be so powerful that they are equivalent with Turing machines. In this paper, a modified inversion particle swarm optimization was proposed, this method and the mutational mechanism of genetics algorithm were used to combine with the tissue-like P system, through these evolutionary algorithms and the P system, the idea of a novel membrane clustering algorithm could come true. Experiments were tested on six data sets, by comparing the clustering quality with the GA-K-means, PSO-K-means and K-means proved the superiority of our method.

Keywords: membrane clustering algorithm, P system, particle swarm optimization, evolutionary algorithm

Copyright © 2016 Institute of Advanced Engineering and Science. All rights reserved.

1. Introduction
Membrane computing is a recent domain of natural computing started by Gh. Păun in 1998 and is known as P systems or membrane systems [1]. It is a parallel distributed computing model inspired from the functionality and structure of living cells and the intersection of them in the tissue, organ and neural network. There are three main classes of P systems investigated: cell-like P systems, tissue-like P systems and neural-like P systems [2], [3, 4]. These computing systems are proved to be so powerful that they are in some ways equivalent with Turing machines [5], the powerful global search capacity of membrane computing is also proved [6].

Cluster analysis also called clustering, is a typical example of unsupervised learning, it aims at partitioning a set of data objects into subsets, each subset called a cluster, such that objects in a cluster are similar to each other, yet dissimilar to objects in other clusters [7, 8]. Clustering analysis has been widely used in various fields, such as image processing, data analysis, Web searching, market analysis etc. As one of the most classical clustering method based on partitioning, K-means algorithm is an unsupervised clustering algorithm for classifying data points into k distinct groups. It is the most classical clustering method due to its simple principle and easily application [9]. However, the final clustering results of it depend heavily on the initial random selection such that the K-means method is not guaranteed to converge to the global optimum and often trapped into the local optimum [10].

As a novel computing model inspired by biology, the P system has attracted more and more people's attention. Due to the parallel distributed computing ability of the P system, a number of papers have appeared in which P system are hybridized with other evolutionary algorithms like Multi-level thresholding methods [11], Particle Swarm Optimization [12, 13], Genetic Algorithms [14], Ant Colony Optimization [15], Quantum Inspired EAs [16], etc. These modifications obviously made many improvements on the evolutionary algorithm and promoted the development of the P system research.

In order to attain a better clustering result, more improvements have made on current clustering methods. In this paper, a modified inversion particle swarm optimization was proposed, this method and the mutational mechanism of genetics algorithm were used to combine with the tissue-like P system, Through the special designed tissue-like P system and the evolution mechanism and communication mechanism, the idea of the novel membrane clustering algorithm could realized. The rest of the paper is organized as follows. Section 2

Received December 28, 2015; Revised April 7, 2016; Accepted May 1, 2016
describes briefly partition-based cluster algorithm. Section 3 introduces tissue-like P system. Section 4 presents a variant of the PSO algorithm and the proposed membrane clustering algorithm. Section 5 illustrates experimental results. Finally, Section 6 makes conclusions.

2. Partition-Based Cluster Algorithm

Cluster analysis is a process of partition data points into different clusters due to their similarity. As a powerful technique of data mining, clustering is widely used in many fields, such as business intelligence, Web searching, biological, safety and so on. Major basic clustering methods can be divided into the following categories: partitioning method, hierarchical method, density-based method, grid-based method [7]. Suppose that dividing the data points into k clusters based on partition-based clustering algorithm, the k cluster centers, $z_1, z_2, \ldots, z_k$, and the corresponding clustering partitions are $C_1, C_2, \ldots, C_k$ respectively. The data point $x_i$ belongs to the cluster $p$ if the distance meets:

$$\|x_i - z_p\| \leq \|x_i - z_q\|, p, q = 1, 2, \ldots, k, p \neq q$$

Once clustering partitions are formatted, new cluster centers are computed by the means of the points in the corresponding cluster using

$$z_p = \frac{1}{n_p} \sum_{x_i \in C_p} x_i$$

Where $n_j$ is the number of data points in cluster $j$ and $C_j$ is the clustering partition of cluster $j$. In this partition-based clustering method, the clustering metric $M$ is determined as follows:

$$M(C_1, C_2, \ldots, C_k) = \sum_{i=1}^{k} \sum_{x_i \in C_i} \|x_i - z_j\|$$

Obviously, smaller $M$ value stands for better clustering results, such that this clustering problem can be considered as an optimization process of finding the smaller $M$.

3. The Tissue-Like P System

Membrane computing is a recent domain of natural computing started by Gh. Păun in 1998. It is also known as membrane computing or membrane systems. In recent years, many different models of P systems have been proposed, such as cell-like P systems, tissue-like P system, and spiking neural P system [17-21]. The obtained computing systems have proved to be so powerful that it is equivalent with Turing machines [22]. The P systems are a class of distributed parallel computing devices of a biochemical type [23], membrane computing now is a hot cross-discipline topic which involves computer science, mathematics, biology and artificial intelligence, etc. One P system mainly consists of three parts: membrane structure, multisets of objects and evolution rules [1]. In tissue-like P system, membranes or cells are placed as the nodes of a graph. The net of nodes deals with symbols and communicates symbols along channels specified in advance. The communication among cells is based on symport/antiport rules [24]. Symport rules move objects across a membrane together in one direction, whereas antiport rules move objects across a membrane in opposite directions [25]. The tissue-like P system which contains $m$ elementary membranes is shown in Figure 1 and described as follows:

$$R = (\omega_1, \ldots, \omega_{m'}, R_1, \ldots, R_m, R', i_0)$$

(1) $\omega_i (1 \leq i \leq m)$ is a finite set of objects in elementary membrane $i$;
(2) $R_i (1 \leq i \leq m)$ is a finite set of evolution rules in elementary membrane $i$;
(3) $R'$ is a finite set of communication rules with the two following forms: antiport rule: $(i, \bar{z}/z', j)$, $i, j = 1, 2, \ldots, m$, $i \neq j$. The rule is used to communicate the objects between an elementary
membrane and its neighbor; symport rule: \((i, Z/\lambda, 0), i = 1, 2, ..., m\). The rule is used to communicate the best objects between membranes and the membrane with the environment. (4) \(b_i\) indicates the output region of the system.

\[
\begin{align*}
\text{Figure 1. The designed tissue-like P system}
\end{align*}
\]

4. The Proposed Membrane Clustering Algorithm

4.1. An Improved Inversion Particle Swarm Optimization

Particle swarm optimization is a population-based stochastic optimal algorithm proposed by Kennedy and Eberhart in 1995. The basic idea of this algorithm simulates bird flocking or fish schooling behavior to build a self-evolving system [18]. As a representative of the swarm intelligence algorithms, particle swarm optimization algorithm is used to solve continuous optimization problems originally [26]. In this algorithm, each particle was considered as a potential solution to an optimal question, these particles flew at a specific speed and then adjust speed with their own and the other particles’ flying experience, every particle has an objective function to determine its fitness which was used for evaluate itself, during this iteration course, the particle flew toward the best position, and the optimization problem tends to obtain its optimal solution[27]. Particle swarm optimization also has some drawbacks, for example, it is easy trap into the local optimum and be absence of population diversity. Some certain modifications have been made in basic PSO algorithm, for instance, inertia weight, reverse searching, parameter modification and so forth [28]. In this paper, we propose a modified inversion particle swarm optimization (MIPSO) based on the idea of avoiding the worst position in order to improve the exploration capacity and the diversity of the whole swarm [29]. Suppose that \(N\) particles exist in \(D\)-dimensional searching space, the particle updates its velocity and position in our method as follows:

\[
\begin{align*}
V_{id}^{k+1} &= \alpha V_{id}^{k} - c_1 r_1 (P_{id}^k - x_{id}^k) - c_2 r_2 (G_{id}^k - x_{id}^k) \\
X_{id}^{k+1} &= X_{id}^k + V_{id}^{k+1}
\end{align*}
\]

where \(d = 1, 2, ..., D; i = 1, 2, ..., N, V_{id}^{k+1}\) is the i-th particle’s new velocity at the kth iteration step, the range of \(V_{id}\) is \([-V_{max}, V_{max}]\), such per-set value prevents the particle from
flying out of the search space; \(^k\) is the iteration number; \(x_{k+1}^i\) is the \(i\)-th particle’s position of the \(k+1\) times, based on its previous position and new velocity at the last iteration; \(\omega\) is inertia weight; \(p_{iw}\) is the worst fitness value of the particle itself achieved so far, and \(p_{gd}\) is the worst global fitness the whole swarm attained so far; \(c_1, c_2\) are learning factors; \(\text{rand}()\) are random numbers uniformly distributed in the range \([0, 1]\). The performance of each particle is evaluated according to a predefined fitness function, which is usually proportional to the cost function associated with the problem. When finding the optimum or meeting the maximal iteration numbers, the searching process halted.

In this method, we consider the particle searching for the best condition by combining the worst flying experience of itself and the worst flying experience of the whole population, the particle adjusts its velocity by flying towards the contrary direction of the worst position, the fitness value is adopted to evaluate the particle. Learning from the worst experience and run away from it is the basis idea of this improved algorithm.

4.2. The Evolution Rules and Communication Rules of the Tissue-Like P System

In each elementary membrane of the tissue-like P system, \(n\) number objects are contained. Every object represents the \(k\) possible cluster centers, evolution rules are applied to evolve the objects, and the evolution mechanism is designed based on particle swarm optimization, its improved version MIPSO and mutation rules of generation algorithm.

In the proposed tissue-like P system, each elementary owns the same \(n\) number objects, during the evolution, these objects are randomly divided into two parts roughly, one of the parts execute standard particle swarm optimization while the other carry out the MIPSO, the two optimize the objects through by finding the best position and avoid the worst position. In this work, mutation mechanism in the binary coding generation algorithm is also used to coevolve the objects, and it makes mutation according to the possible \(pm\), if the value of a mutation point at dimension \(j\) is \(v\), after mutation the value becomes that

\[
v' = \begin{cases} 
  v \pm \delta, & v \neq 0 \\
  v, & v = 0 
\end{cases}
\] (7)

In the above formula, the signs “+” or “−” occur with equal probability, and \(\delta\) is real number generated with uniform distribution in the range \([0, 1]\). The optimization process halted until it meets the maximum iteration criteria.

The communication rules in the designed P system mainly have two types: antiport rule: \((i, Z, i', j), \ i, j = 1, 2, \ldots, m, \ i \neq j,\) including \(m\) is the number of elementary membranes, this rule is used to communicate the objects between an elementary membrane and its two neighbor; symport rule: \((i, Z, 0), \ i = 1, 2, \ldots, m,\) this rule is used to communicate the objects between membrane and the environment, of which \(m\) is the number of the elementary membranes. The label 0 stands for the output region namely environment. The role of the communication mechanism is to communicate the optimal object of each membrane and eventually attain the global optimum which stores in the output region. The reference point is the fitness objection that is the \(M\) value.

4.3. Description of the Proposed Membrane Clustering Algorithm

In this paper, a new membrane clustering algorithm based on partitioning clustering is proposed through utilizing the rules of evolutionary algorithm, communication rules and the special framework within the distributed parallel computation framework provided by tissue-like P system. The objects of each cell are represented as the possible cluster centers. Particle swarm optimization algorithm and a modified reverse particle swarm optimization algorithm are applied as evolution mechanism to evolve the objects. Mutation mechanism of the generation algorithm is also imported as coevolution rules in order to improve the diversity of the population. Communication rules between cells or between cells and environment is used to exchange the object and arrive to attain the optimum. This new membrane clustering algorithm can automatic searching the optimum cluster centers and arrive at better cluster results, the clustering algorithm is described as follows:
Step 1. Generate \( n \) initial objects (the possible cluster centers) for each of the \( m \) elementary membrane, then distinguish clustering regions according to the existing formula and calculate the new cluster centers.

Step 2. Initial a group of individuals randomly, divide the individuals into two portions with the equal scale, initial the velocity and the position, set the mutation probability \( p_m \), fitness function \( f \).

Step 3. Execute the particle swarm optimization and the modified inversion particle swarm optimization, mutation rules are applied simultaneously.

Step 4. Communication rules are used to exchange the optimum between the elementary membranes and elementary membrane with environment.

Step 5. Repeat these steps until attain the maximum iteration criteria, and the final global optimum is showed in the environment.

5. Experiments and Results

The proposed membrane clustering algorithm is assessed on six data sets and compared with K-means algorithm, PSO-K-means algorithm and GA-K-means algorithm [30, 31]. For illustrating the results of the proposed algorithm exactly, 30 runs were executed when applying one of the tested algorithms. These six data sets including two artificial data sets named Art1, Art2 and four real life data sets named Iris, Cancer, Vowel, Wine are used. All data sets except Art1 and Art2 are available at ftp://ftp.ics.uci.edu/pub/machine-learning-databases/.

Table 1 summarizes the characteristics of these data sets. Data sets Art 1, Art 2 is illustrate in Figure 2, Figure 3.

<table>
<thead>
<tr>
<th>Name of Date set</th>
<th>NO. of classes</th>
<th>NO. of features</th>
<th>Size of data set (size of classes in parentheses)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art1</td>
<td>3</td>
<td>3</td>
<td>900 (300, 300, 300)</td>
</tr>
<tr>
<td>Art2</td>
<td>2</td>
<td>3</td>
<td>900 (450, 450)</td>
</tr>
<tr>
<td>Iris</td>
<td>3</td>
<td>4</td>
<td>150 (50, 50, 50)</td>
</tr>
<tr>
<td>Cancer</td>
<td>2</td>
<td>9</td>
<td>683 (444, 239)</td>
</tr>
<tr>
<td>Vowel</td>
<td>6</td>
<td>3</td>
<td>871 (72, 89, 172, 151, 207, 180)</td>
</tr>
<tr>
<td>Wine</td>
<td>6</td>
<td>13</td>
<td>178 (59, 71, 48)</td>
</tr>
</tbody>
</table>

Figure 2. Art 1
The values of the parameter in these tested algorithms are shown in Table 2.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\omega$</th>
<th>$c_1$</th>
<th>$c_2$</th>
<th>$P_m$</th>
<th>M</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>0.9~0.4</td>
<td>1.2</td>
<td>1.2</td>
<td>0.05</td>
<td>1000</td>
</tr>
</tbody>
</table>

$\omega$ is the inertia weight, the value is between 0.4 and 0.9, small value improves local convergence capacity and the large value improves global convergence capacity; $c_1$ and $c_2$ are learning factor; $P_m$ is mutation probability; M is the maximum iteration. In order to investigate the effect of elementary membrane numbers, four different tissue-like P system with 4, 8, 16, 20 elementary membranes are designed. In order to overcome the effects of accidental factors, we choose the average value of the every whole 30 times runs. The M values of these P systems are shown in Table 3.

<table>
<thead>
<tr>
<th>Data set</th>
<th>Art 1</th>
<th>Art 2</th>
<th>Iris</th>
<th>Cancer</th>
<th>Vowel</th>
<th>Wine</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art 1</td>
<td>628.3298</td>
<td>627.2565</td>
<td>624.4532</td>
<td>634.0342</td>
<td>624.4532</td>
<td></td>
</tr>
<tr>
<td>Art 2</td>
<td>546.7652</td>
<td>545.3545</td>
<td>539.2868</td>
<td>539.7662</td>
<td>539.7662</td>
<td></td>
</tr>
<tr>
<td>Iris</td>
<td>99.9238</td>
<td>98.7596</td>
<td>98.6324</td>
<td>98.9328</td>
<td>98.9328</td>
<td></td>
</tr>
<tr>
<td>Cancer</td>
<td>3253.3342</td>
<td>3052.5372</td>
<td>3249.6575</td>
<td>3250.0442</td>
<td>3250.0442</td>
<td></td>
</tr>
<tr>
<td>Vowel</td>
<td>149315.4096</td>
<td>149311.8741</td>
<td>149309.4863</td>
<td>149309.8223</td>
<td>149309.8223</td>
<td></td>
</tr>
<tr>
<td>Wine</td>
<td>16313.6521</td>
<td>16306.4536</td>
<td>16293.5662</td>
<td>16293.7976</td>
<td>16293.7976</td>
<td></td>
</tr>
</tbody>
</table>

Table 3 indicates that tissue-like P system with 16 elementary membranes behaves better than the others with the smaller M value in these six test data sets. For the true life data set of Cancer, the performance of the 16 elementary membrane P system is 3249.6575, superior to the other 3251.3342, 3052.5372 and 3250.0442 obviously.

Generally, more particles in the swarm could enhancing the searching range and may leads better optimization results. To inquiry the appropriate number of the particles, different scale of the swarms are designed for the experiment. Table 4 shows the diverse results of the different particle scales.

<table>
<thead>
<tr>
<th>Data set</th>
<th>20</th>
<th>50</th>
<th>100</th>
<th>200</th>
<th>400</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art 1</td>
<td>629.2378</td>
<td>624.9662</td>
<td>624.5342</td>
<td>624.4532</td>
<td>624.4532</td>
</tr>
<tr>
<td>Art 2</td>
<td>550.4574</td>
<td>546.6753</td>
<td>539.7662</td>
<td>539.2868</td>
<td>539.2868</td>
</tr>
<tr>
<td>Iris</td>
<td>104.2438</td>
<td>101.7456</td>
<td>99.9458</td>
<td>98.6324</td>
<td>98.6324</td>
</tr>
<tr>
<td>Cancer</td>
<td>3257.5664</td>
<td>3053.8698</td>
<td>3252.6572</td>
<td>3249.6575</td>
<td>3249.6575</td>
</tr>
<tr>
<td>Vowel</td>
<td>149320.4033</td>
<td>149314.6541</td>
<td>149310.6222</td>
<td>149309.4863</td>
<td>149309.4863</td>
</tr>
<tr>
<td>Wine</td>
<td>16317.3496</td>
<td>16311.4581</td>
<td>16295.2803</td>
<td>16293.5662</td>
<td>16293.5662</td>
</tr>
</tbody>
</table>
Table 4 shows that when the number of particles attains 200, the clustering quality will be better than the small-scale swarms. For the Wine data set, the M value is 16293.5662 with the particle scale is 200, obviously smaller than the smaller scale with the value 16317.3496, 16311.4581, 16295.2803. Too many particles may do no help in improving the clustering quality, the clustering results have no change when the number rose from 200 to 400, and thus the appropriate number could define 200 in our membrane clustering method.

<table>
<thead>
<tr>
<th>Data set</th>
<th>GA</th>
<th>PSO</th>
<th>K-means</th>
<th>MCA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Art 1</td>
<td>649.9533</td>
<td>639.2434</td>
<td>649.1573</td>
<td>623.3662</td>
</tr>
<tr>
<td>Art 2</td>
<td>542.0232</td>
<td>541.6982</td>
<td>552.3472</td>
<td>539.0848</td>
</tr>
<tr>
<td>Iris</td>
<td>98.9734</td>
<td>98.0531</td>
<td>106.1347</td>
<td>97.0324</td>
</tr>
<tr>
<td>Cancer</td>
<td>3249.8713</td>
<td>3048.2311</td>
<td>3257.6725</td>
<td>3248.5472</td>
</tr>
<tr>
<td>Vowel</td>
<td>149310.0908</td>
<td>149309.7492</td>
<td>149336.9646</td>
<td>149307.4223</td>
</tr>
<tr>
<td>Wine</td>
<td>16293.3142</td>
<td>16292.0419</td>
<td>16318.4763</td>
<td>16290.7976</td>
</tr>
</tbody>
</table>

According to the results of the chart in Table 5, we could conclude that, in most of runs of these algorithms, the proposed membrane clustering method is superior to other tested algorithm. For Art 1, the value of the MCA is 623.3662, much less than the PSO which is 639.2334, the GA which is 649.9533 and the K-means algorithm which is 649.1573. The M value on the Vowel is 149307.4223, smaller than the K-means, GA-K-means and PSO-K-means with 149336.9645, 149309.7492 and 149310.0908, respectively. 30 times runs of each tested algorithm lows the contingency of the experiment and enhance the persuasion of the test. Through comparing and analyzing, the better clustering quality of the proposed membrane clustering algorithm could be proved.

6. Conclusions

At present there are many novel techniques committed to improving the data clustering. This paper focuses on realizing a membrane clustering algorithm based on a designed tissue-like P system. With the structure of the tissue-like P system, particle swarm optimization and an modified inversion particle swarm optimization are applied as the evolution rules and the mutation mechanism of the generation algorithm is used as the supplement of the evolution rules and enhance the diversity of the population simultaneously, communication rules which contain antiport rules and symport rules are employed to communicate the optimum, our proposed membrane clustering algorithm behaves superior to the some other clustering algorithm, the perfect performance result of the simulation experiment prove this claim. The further work concentrate upon using more evolutionary algorithm based on membrane computing in order to improve the clustering methods and enhance the clustering quality.

Acknowledgements

Project supported by National Natural Science Foundation of China (61472231, 61170038, 61502283), Jinan City independent innovation plan project in College and Universities, China (201401202), Ministry of education of Humanities and social science research project, China (12YJA630152), Social Science Fund Project of Shandong Province, China (11CGLJ22).

References


