A PSO-based Clustering Algorithm Inspired by tCP System

Tong Gao [a], Xiyu Liu [b,*]

College of Management Science and Engineering, Shandong Normal University, Jinan, Shandong, 250014. China

[a] gtsdnu@163.com, [b] sdxyliu@163.com

Abstract

Particle swarm optimization (PSO) is a biological imitation algorithm, which has become a basic and commonly used optimization algorithm. It has many advantages, such as high precision and fast convergence, which attracts the academic attention. P systems, also known as Membrane systems, is a class of distributed parallel computing models. This paper presents a PSO-based clustering algorithm inspired by a doubly-linked P system with chain structure. The proposed PSO-based clustering algorithm adopts the structure of neighborhood, in which population evolution is guided by neighborhood information. In addition, a variable size neighborhood structure is used to control the particle swarm optimization and convergence. Adopting the doubly-linked P system with chain structure(tCP), the information exchange among multiple populations is realized, and the population diversity is increased. The characteristics of the distributed parallel computing of the membrane system can accelerate the convergence of the population Experimental results show that the proposed algorithm outperforms several evolutionary clustering algorithms recently reported, such as GA, DE, PSO and K-means algorithm.

Keywords

P system, Clustering, PSO, Evolutionary algorithms.

1. Introduction

Particle swarm algorithm is a random search algorithm, to optimize the various functions effectively. It is committed to providing a solution to the optimization problems of complex systems, including data clustering. As a data mining techniques, data clustering aims to find out the most natural partition of a data set such that data points in the same cluster are as similar as possible to each other while data points from different clusters share the minimum similarities. Up to now, numerous algorithms have been raised. However, it is a long process to improve the performance of the optimization algorithm. In view of the above issue, a PSO-based clustering algorithm inspired by doubly-linked P system with chain structure, is proposed in this paper.

P system, also named Membrane systems, as a new class of distributed parallel computing models, is inspired by the structure and function of biological cells and also the tissues, organs and populations of cells [14]. Scholars have proposed a variety of membrane systems: Tissue-like P systems with channel states [5], Spiking neural P systems [6], Tissue-like P Systems with Protein on Cells [15], etc. In this paper, The proposed tCP consists of four cells linked as a chain, where each cell maintains a population of particles. Two-way communication rules are used between adjacent cells, where the best objects in each cell will be transported into its adjacent cells in iterations. Meanwhile, particle swarm in each cell adopts different scale neighborhood structure to realize particle evolution, which enhances the diversity of objects in the system.

The remainder of this paper is arranged as follows. Section 2 gives a theoretical description of presented tCP system. And an improved PSO algorithm based on neighborhood is proposed, too. The detailed description of the proposed membrane clustering algorithm is given in Section 3. Experimental setting and results of test on six data sets, as well as analyses are provided in Section 4, followed by conclusions in Section 5.

2. Preliminary

Data clustering amounts to find out the most natural partition of a data set such that data points in the same cluster are as similar as possible to each other while data points from different clusters share the minimum similarities. By far, a variety of intelligent optimization methods have been proposed to solve data clustering problems. The proposed PSO-based clustering algorithm combines a novel P system and an improved particle swarm algorithm, named FERPSO.

2.1 Doubly-linked P system with chain structure

A variety of P system models have been proposed. Liu et al. proposed an improved apriori algorithm based on an evolution-communication tissue-like P system with promoters and inhibitors in [11]. A tissue-like P system with a loop topology structure of cells was presented in [13] to solve clustering problem. Considering each cell as a node, the communication rules establish a virtual graph, where the edges connecting two cells indicate a executable communication rule. With different structures and rules, P systems are able to provide impactful solutions for practical problems. P system with chain structure is a distributed parallel computability model, based on the notion of a membrane structure. Such a structure consists of several single-membrane cells, recurrently placed inside a common environment [17]. This paper adopts the two-way communication rules of P system, which contains 4 cells linked as a chain. The proposed tCP system is shown in 0.

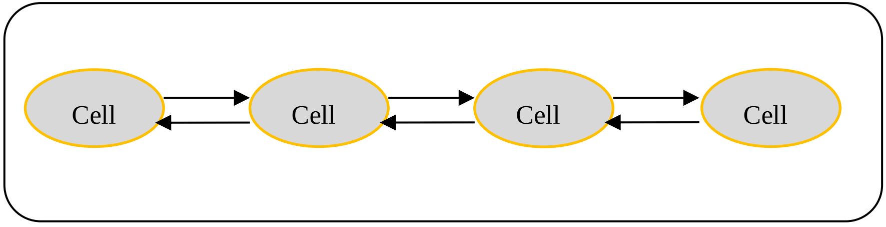

Fig.1.Doubly-linked P system with chain structure

2.2 Neighbors-guide particle swarm algorithm

The same as other evolutionary computation algorithms, Particle Swarm Optimization (PSO) [8] is a population-based stochastic search process that is first introduced by Kennedy and Eberhart. Every particle represent a potential solution to a specific fitness function. The target of PSO is to search a set of particles to optimize the fitness function. By evolving the particles according to a specified velocity update formula iteratively, PSO gradually achieves the purpose. In classical PSO, the particle's velocity continuously updates according to it's best previous success, the best success of one neighborhood particle, the particle's current position, and its current velocity.

However, important information contained in other neighborhood particles is neglected through overemphasis on the single global best. And that may leads to limited search capability or slower convergence. A particle swarm optimizer based on fitness Euclidean-distance ratio[9], termed PERPSO, is proposed to handle the issue. In FERPSO, both of the best success of current particle, but also the neighborhood best to each particle is used to guide the evolution of particles. To be specific, a particle moves toward its personal best as well as its fitness-closest neighbors, which are indicated by the fitness-Euclidean distance ratio (FER) values. The nBest for i-th particle is selected as the neighborhood personal best with the largest FER as follows:

$$FER_{(i,j)} = \alpha \cdot \frac{f(p_j) - f(p_i)}{\|p_j - p_i\|}$$

Where $\alpha = \frac{\|s\|}{f(p_g) - f(p_w)}$ is a scaling factor, p_w is the worst-fit particle in the current population, p_i and p_j represent the personal bests of the i-th and j-th particle, respectively, and s is the size of the search space, which is estimated by its diagonal distance $\sqrt{\sum_{k=1}^{d}(x_k^u - x_k^l)^2}$ (where x_k^u and x_k^l are the upper and lower bounds of the k-th dimension of the search space).

How to use the information of neighbors determines how diverse the influence will be and how efficient the algorithm will be. Different with the original version of FERPSO, after getting the fitness-closest neighbors, nBest for i-th particle is generated by the following equation:

$$nBest = \frac{\sum_a^{ns} nBest_a}{n}$$

where $nBest_a$ denotes the a-th fitness-closest neighbor of particle i and ns is neighborhood size(number of neighbors). The global best solution $gBest$ in original velocity update equation is replaced by $nBest$. Specifically, the velocity update equation is rewritten as:

$$V_i = \omega V_i + c_1 r_1 (pBest_i - X_i) + c_2 r_2 (nBest - X_i) \tag{1}$$

In the above equation, V_i, X_i indicate the velocity and position of particle i, respectively. And ω denotes the inertia weight, r_1, r_2 are randomly sampled from a uniform distribution in the range $[0, 1]$, c_1, c_2 denotes the acceleration constants.

3. The PSO-based clustering algorithm inspired by Doubly-linked P system with chain structure

In this section, we give a detailed description of the PSO-based clustering algorithm inspired by doubly-linked P system with chain structure, in which cluster centers are determined as the local maxima of the fitness function defined in Section 3.2. This fitness function characterizes the compactness of a potential solution. The higher the compactness is, the smaller the fitness value tends to be, the more probably this potential solution is to be a global optimum solution, as well. As for the evolution of the objects, we adopted equation (1), in which nBest is calculated by more than one fitness-closest neighbors. Details can be seen in Section 3.3. To realize the communication between the cells, the best object in each cell are transported to the adjacent cells by an two-way communication rules proposed in section 3.4. The process of proposed clustering algorithm is described in Algorithm1.

3.1 Initialization

At the beginning of algorithm, location matrix X of population in each cell and its velocity matrix V are to be initialized. The population size N_p in each cell is set to 50 and the value of the maximum iterations T = 200. Acceleration constants c_1 and c_2 are set to 2.05. The inertia weight ω is 0.7298. In addition, the numbers of neighbor particles ns in four cells are set to be 2-5 respectively.

3.2 Fitness Function

Data clustering is a process of dividing data points into clusters such that data points in the same cluster are as similar as possible. For a data set X consists of N data points, assume that a partition is made up of k clusters $C = \{C_1, C_2, \cdots, C_k\}$ with cluster centers $\{m_1, m_2, \cdots, m_k\}$ is given, where each cluster C_i represents a subset of data points. A widely used fitness function , the within-cluster error $\sum_{i=1}^{k} \sum_{X_j \in C_i} \left\| X_j - m_i \right\|^2$ is transformed to evaluate the evolution of particle swarm. For a particle u (candidate solution), its fitness we adopted is defined as:

$$fitness(u) = \sum_{i=1}^{k} \sum_{X_j \in C_i} \left\| X_j - m_i \right\|$$

A smaller fitness value means a better partition. That is to say, a particle with smaller fitness value indicates a more competitive solution. We attempt to find a set of cluster centers $\{m_1, m_2, \cdots, m_N\}$ which has the minimum value of fitness function.

3.3 Evolution of Objects

The evolution rules are used to evolve the objects associated with cluster centers in cells, thus the algorithm is able to find the optimal cluster centers for a dataset by means of the evolution of objects.

Algorithm1. PSO-based Clustering Algorithm Inspired by tCP System

```
Input: $X$ (Data set), $k$ (the number of clusters), $N_P$ (the number of objects in each cell), $T$ (maximum iterations), $Ns$ (the number of  neighbor particles)
Output:  $G$ (the optimal centers)
Initialize the objects in cells
$t \leftarrow 0$
while ($t < T$) do
    For each cell $c$ in parallel do
        Fill up its population with the better object from two adjacent cells
        Evolve its objects by using improved PSO algorithm
        Transport its best $O_b$ objects into next cell by communication rule of type Ⅰ
    End
    Update the global best object of the system by communication rule of type Ⅱ
    $t \leftarrow t + 1$
End
Partition $N$ data points into the Corresponding clusters
```

After getting the objects from its adjacent neighbor cells, the objects evolve by using improved PSO algorithm. What's different with FERPSO is that the evolution of objects takes different neighborhood structures. That is to say, in each cell, the velocities of objects update according to the equation with different ns values described in section 2.2. After that, all particles fly to the new position and the fitness of each object is calculated according to fitness function mentioned in section3.2. By sorting the fitness value of objects, the best object in each cell is selected to be transported into the adjacent neighbor cells as well the objects from adjacent neighbor cells are compared. Then the better one of two objects from adjacent neighbor is picked out to participation evolution in the next iteration.

3.4 Communication between Cells

The designed tCP system adopts the communication rules of two types:

(Ⅰ) Communication rules used to transport the better objects between adjacent cells.

(Ⅱ) Communication rules used to transport the best object in each cell into environment to chose global best object.

The four cells in the tCP system establish the annular communication relationship of objects by first communication rule(Ⅰ) (seen in 0). For each cell, the best object is transported into its neighbor cells and one of the objects communicated from its adjacent cells will be singled out to fill up the population for next computing step. Meanwhile, each cell communicates its best object $CBest$ into the environment by using second communication rule(Ⅱ) and global best object $gBest$ is selected from the $CBest$. Note that in each computing step, the communication rules are executed after the evolution rules in maximum parallel way under the control of a global clock.

4. Experimental results

In terms of experiment, the performance of proposed clustering algorithm is tested on a synthetic data set [1] and three real-life data sets [10], including Iris, Glass and Wine. In what follows, a description of experimental settings is first given. Then, the performance comparison between proposed clustering algorithm and four popular clustering algorithms is revealed.

4.1 Experimental setting

In this section, we compare the proposed PSO-based clustering algorithm inspired by tCP system with K-means [4], PSO [7], and another two evolutionary clustering algorithms: GA [2] and DE [16]. These algorithms are implemented in Matlab2016a according to the following parameters.

- PSO. The inertia weight ω uses a linear decreasing, where initial $\omega = 1$ and damping ratio $\omega_d = 0.98$, the population size N_p = 50, acceleration constants c_1 and c_2 are set to 2, and maximum iteration number is 200.
- GA. The crossover and mutation probabilities, p_c and p_m, are chosen to be 0.4 and 0.2, respectively. Let the population size be N_p = 50 and let maximum iteration number be T = 200.
- DE. Crossover rate CR = 0.2 while the scaling factor is randomly generated in [0.4, 0.8]. Equally, let the population size be N_p = 50 and the maximum iteration number is set to be T = 200.
- K-means. The maximum iteration number is set to be T = 200.

In the experiments, both of synthetic and real-life data sets are used to evaluate these clustering algorithms. The manually generated data set Data_4_3 arise from the existing literatures. More details of all data sets are briefly described in 0.

Table 1.Details of data sets.

Dataset	Source	Data points	Dimension	Clusters
Data_4_3	synthetic	400	3	4
Iris	UCI	150	4	3
Glass	UCI	214	9	6
Wine	UCI	178	13	3

For each data set, every clustering algorithms are executed 20 times repeatedly to ensure the accuracy of the results during the experiments.

4.2 Results and analysis

0 gives the comparison results of four clustering algorithms on the six data sets, respectively. The experimental results reveal that the proposed algorithm provides the optimum average value compared to the other four clustering algorithms. As we see in 0, the results obtained on the Data_4_3 indicate that the proposed algorithm converges to the optimum of 801.79 while PSO, GA, DE and K-means attain 826.32, 820.41, 937.89 and 1383.40 respectively. It's also obvious that the experiments on the Glass provide a distinction where the optimum value of proposed algorithm is 280.58 while the PSO, GA, DE, and K-means obtain 289.19, 306.61, 293.53, and 470.93, respectively. Compared to several existing clustering algorithms, the proposed clustering algorithm outperforms remarkably. Furthermore, test results on other data sets also prove the better effectiveness of the proposed algorithm.

Table 2. Mean results of five algorithms running on 6 data sets over 50 times.

Data sets	tCP	PSO	GA	DE	K-means
Data_4_3	**801.79**	826.32	820.41	937.89	1383.40
Iris	96.69	98.66	**96.42**	98.00	162.28
Glass	**280.58**	289.19	306.61	293.53	470.93
Wine	**16313.69**	16360.10	16314.14	16319.62	19518.90

To demonstrate the superiority of the proposed algorithm, we give a more intuitive presentation about the particular performance of different clustering algorithms in 0. The blue line represented the proposed algorithm converge more faster in the early iterations than PSO, GA, DE and K-means as shown in 0. All of the results obtained from four datasets are convincing evidences of the proposed clustering algorithm converging faster.

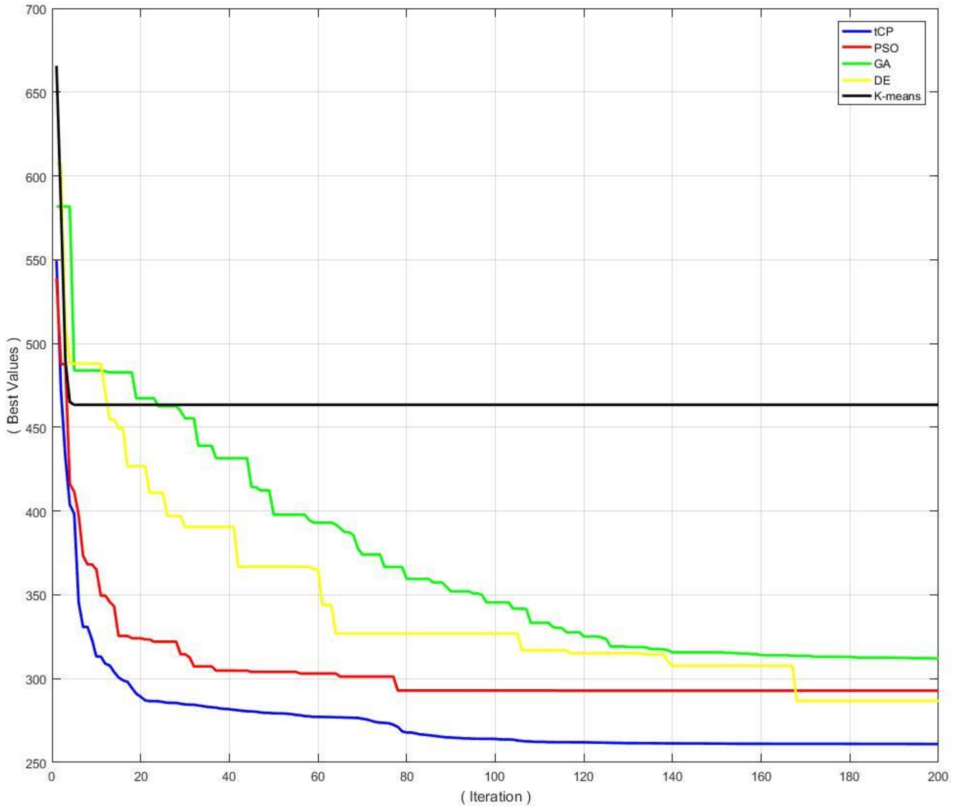

Fig.2.The running results of proposed algorithm on data sets Glass

It is worth mentioning that the results on the real data set Iris, the optimum value of proposed algorithm is 96.65, which is higher than the results of GA. However, the proposed clustering algorithm obtains smallest standard deviation of fitness in comparison to the other four algorithms, as shown in 0. After running of each algorithm on data set Iris independently for 30 times, the variance obtained by proposed algorithm is 0.0285. As for PSO, GA, DE and K-means, they attain 0.3898, 7.7226, 2.7825, 385.1923 respectively. Obviously, the variance of proposed algorithm is much smaller than all other four algorithms. Therefore we insist that the proposed clustering algorithm is much more robust than the other four algorithms.

Table 3. Statistical analysis of results running on data set Iris over 30 times.

Algorithm	tCP	PSO	GA	DE	K-means
mean	96.6943	98.6636	96.4240	101.6337	170.9794
variance	**0.0285**	0.3898	7.7226	2.7825	385.1923
minimum	96.6555	97.6678	83.7182	98.0040	162.2799
maximum	97.5796	99.8327	99.0224	104.8482	214.4015

It is worth mentioning that the results on the real data set Iris, the optimum value of proposed algorithm is 96.65, which is higher than the results of GA. However, the proposed clustering algorithm obtains smallest standard deviation of fitness in comparison to the other four algorithms, as shown in 0. After running of each algorithm on data set Iris independently for 30 times, the variance obtained by proposed algorithm is 0.0285. As for PSO, GA, DE and K-means, they attain 0.3898, 7.7226, 2.7825, 385.1923 respectively. Obviously, the variance of proposed algorithm is much smaller than all other four algorithms. Therefore we insist that the proposed clustering algorithm is much more robust than the other four algorithms.

5. Conclusion

In this paper, we have proposed a PSO-based clustering algorithm inspired by doubly-linked P system with chain structure. Each cell, as a parallel computing unit in designed tCP system, runs in maximally parallel way and each object of the system denotes a group of candidate centers. Two kinds of rules are adopted in the proposed algorithm: communication rules and evolution rules. The communication rules build a local neighborhood topology in virtue of the chain structure of cells, where the best objects are transported to adjacent cells. Distinguished from the existing evolutionary algorithm, improved PSO-based evolution rules are used to evolve objects, which is beneficial to accelerate convergence of population. Moreover, test results on data sets demonstrate the superiority of the proposed PSO-based clustering algorithm inspired by tCP system compared to several evolutionary clustering algorithms recently reported.

References

[1] Bandyopadhyay S, Pal S K. Classification and learning using genetic algorithms: applications in bioinformatics and web intelligence[M]. Springer Science & Business Media, 2007.

[2] Bandyopadhyay S, Maulik U. An evolutionary technique based on K-means algorithm for optimal clustering in RN[J]. Information Sciences, 2002, 146(1): 221-237.

[3] B. Qu, A distance-based locally informed particle swarm model for multimodal optimization, IEEE Trans. Evol. Comput. 17(3) (2013)387–402.

[4] Forgy E W. Cluster analysis of multivariate data: efficiency versus interpretability models[J]. Biometrics, 1965, 61(3): 768-769.

[5] Freund R, Păun G, Pérez-Jiménez M J. Tissue P systems with channel states[J]. Theoretical Computer Science, 2005, 330(1): 101-116.

[6] Ionescu M, Păun G, Yokomori T. Spiking neural P systems[J]. Fundamenta informaticae, 2006, 71(2, 3): 279-308.

[7] Kao Y T, Zahara E, Kao I W. A hybridized approach to data clustering[J]. Expert Systems with Applications, 2008, 34(3): 1754-1762.

[8] Kennedy J. Particle swarm optimization[M]//Encyclopedia of machine learning. Springer US, 2011: 760-766.

[9] Li X. A multimodal particle swarm optimizer based on fitness Euclidean-distance ratio[C]//Proceedings of the 9th annual conference on Genetic and evolutionary computation. ACM, 2007: 78-85.

[10] Lichman, M. (2013). UCI Machine Learning Repository, http://archive.ics.uci.edu/ml. Irvine, CA: University of California, School of Information and Computer Science.

[11] Liu X, Zhao Y, Sun M. An Improved Apriori Algorithm Based on an Evolution-Communication Tissue-Like P System with Promoters and Inhibitors[J]. Discrete Dynamics in Nature and Society, 2017, 2017.

[12] Pal N R, Bezdek J C. On cluster validity for the fuzzy c-means model[J]. IEEE Transactions on Fuzzy systems, 1995, 3(3): 370-379.

[13] Peng H, Luo X, Gao Z, et al. A novel clustering algorithm inspired by membrane computing[J]. The Scientific World Journal, 2015, 2015.

[14] Pérez-Jiménez M J, Riscos-Núnez A, Romero-Jiménez A, et al. Complexity-membrane division, membrane creation[J]. The Oxford Handbook of Membrane Computing, 2010: 302-336.

[15] Song B, Pan L, Pérez-Jiménez M J. Tissue P Systems with Protein on Cells[J]. Fundamenta Informaticae, 2015, 144(1):77-107.

[16] Storn R, Price K. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces[J]. Journal of global optimization, 1997, 11(4): 341-359.

[17] Xue J, Liu X. Communication P System with Oriented Chain Membrane Structures and Applications in Graph Clustering [J]. Journal of Computational & Theoretical Nanoscience, 2016, 13(7):4198-4210.