IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received September 9, 2019, accepted September 26, 2019, date of publication September 30, 2019,
date of current version October 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2944675

Hybrid Chain-Hypergraph P Systems for Multiobjective Ensemble Clustering

SHUO YAN, YUAN WANG, DETING KONG, JINYAN HU, JIANHUA QU, XIYU LIU, AND JIE XUE, (Member, IEEE)
Business School, Shandong Normal University, Jinan 250014, China

Corresponding authors: Xiyu Liu (xyliu@sdnu.edu.cn) and Jie Xue (jiexue@sdnu.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61802234 and Grant 61876101, in part by the Natural Science Foundation of Shandong Province under Grant ZR2019QF007, and in part by the China Postdoctoral Project under Grant 2017M612339.

ABSTRACT Clustering is a classic combined optimization problem that is widely used in pattern recognition, image processing, market analysis and so on. However, the efficiency of clustering algorithms decreases as the amount of data increases. In addition, most of the existing methods optimize only one objective and therefore may be suitable only for datasets with certain features. To address these limitations, in this paper, we develop a new hybrid chain-hypergraph P system (named HCHPS), which makes full use of the parallelism of P systems as well as the advantages of chain and hypergraph topology structures for accurate and efficient clustering. Our new P system comprises three types of subsystems, i.e., reaction chain membrane subsystems, local communication membrane subsystems and global ensemble membrane subsystems. Each type of subsystems is implemented end-to-end in HCHPS with new rules and membrane structures in parallel. In particular, to obtain efficient clustering center objects and make the algorithm robust to data with various features, the reaction chain membrane subsystems perform three different multiobjective strategies simultaneously by new chain evolution rules. To increase the population diversity of cluster centers, the local communication membrane subsystems utilize transport rules between membranes for coevolution of nondominated objects. The global ensemble membrane subsystems conduct a new dense representation multisize ensemble strategy to further improve the accuracy of the final results. Evaluations on two artificial data sets and 17 real-life data sets demonstrate the robustness of the proposed method in correctly clustering data sets with different dimensions and shapes. Our experimental results outperform those of both baseline and state-of-the-art methods. Moreover, benefiting from the parallelism, HCHPS is less time consuming than other methods, featuring an average completion time of 28.07 seconds on the 17 real-life data sets. Moreover, an ablation study shows that our proposed components are critical for effective cluster analysis.

INDEX TERMS Chain-hypergraph P system, multiobjective optimization, cluster analysis.

I. INTRODUCTION

Clustering is regarded as a complex optimization problem that plays an important role in data mining and is widely used in pattern recognition, image processing, market analysis and so on. However, the performance of clustering depends on many parameters [1]. Most of the existing methods focus on only one single objective, which cannot achieve optimal results [2]–[4]. To solve this problem, the multiobjective optimization algorithm has been applied in cluster analysis, which can optimize different objective functions simultaneously and help improve the clustering performance [5]–[7]. Saha *et al.* [5], [6] presented a new multiobjective clustering method with two objective functions being optimized together. They also adopted a simulated annealing-based method in multiobjective clustering as a new optimization strategy [7]. The results showed that these multiobjective optimization algorithms improved the performance of cluster analysis.

Different algorithms focus on various features of data sets, and the clustering results are often complementary. To utilize these complementarities in multiple algorithms and improve the accuracy of clustering, ensemble learning has been introduced into cluster analysis [8]–[10]. Iam-On *et al.* [8] presented a new link-based approach to improve the ensemble-information matrix. Wang [9] proposed the

The associate editor coordinating the review of this manuscript and approving it for publication was Shubhajit Roy Chowdhury.

CA-tree to facilitate efficient and scalable cluster ensembles for coassociation matrix-based algorithms. Huang *et al.* [10] proposed an ensemble clustering approach based on sparse graph representation and probability trajectory analysis.

Membrane computing model (also called a P system/membrane system), initiated by Păun [11], is a computational model that encapsulates the data in arrangements of "membranes" that communicate under certain rules with a given computational purpose. Membrane computing has been applied in various fields, such as language generation, electricity fault diagnosis, and image processing [12]–[15]. Clustering based on membrane systems has shown good convergence, robustness, and parallelism [16]–[19]. Peng *et al.* [16] proposed a tissue-like P system based multiobjective fuzzy clustering algorithm to optimize three objectives simultaneously. Qin *et al.* [17] proposed a hybrid clustering algorithm based on a P system and immune mechanism. Peng *et al.* [18] proposed a novel automatic fuzzy clustering method based on an extended membrane system with active membranes. Gao *et al.* [19] presented an improved PSO-based clustering algorithm inspired by tissue-like P system, called TPCA. Besides, new P systems are also designed to solve more problems. Peng *et al.* [20]–[22] proposed a new kind of neural-like P systems by adding threshold, synaptic channels and coupled mechanism. Song *et al.* [23] introduced a new variant of spiking Neural P systems with learning functions to recognize English letters. However, the methods described above used classical P systems. For computation purposes, the classical membrane systems are a simplification of real membrane structures and do not use the complex membrane structures to solve problems with complicated structures; for example, classical P systems cannot store multivariate data with complex relationships. Therefore, there is a pressing need to develop new P systems with complex structures to deal with more real applications. Liu and Xue [24] first attempted to establish P systems on simple complex of discrete objects. Luan and Liu [25] designed P systems on a chain, which can use the directionality of membranes and the additivity of a chain to implement crossover operation and variation operation in genetic or differential evolution algorithm.

In addition, hypergraph theory [26], [27] has been used in clustering, association rule mining, spatial data mining and so on. In hypergraph theory, objects with common attributes belong to one set, and different abstract levels belong to supersets, which comprise a special logic structure that can be used to organize complex relationships between objects. Therefore, making use of the complex structure of hypergraphs in membranes may improve the performance of membrane systems in real applications.

Based on the above considerations, we propose hybrid chain-hypergraph P systems (HCHPSs) to implement multiobjective ensemble clustering. We designed three new types of subsystems with new rules and membrane structures to implement multiobjective optimization, increase the population diversity of cluster centers and conduct an ensemble strategy. The average F-measure on 8 University of California Irvine (UCI) datasets [28], i.e., Iris, Newthyroid, Wine, Diabetes, Bupa, Yeast, Glass, and Cancer are 1.00, 0.92, 0.95, 0.79, 0.78, 0.60, 0.54, 0.97, respectively, outperforming the results of state-of-the-art methods.

The contributions of our work can be summarized as follows:

(1) We propose a new P system with hybrid structures, where the hybrid structures combine the advantages of chain and hypergraph topology structures. We also design new rules in the new P system to solve complex real applications.

(2) A reaction chain membrane subsystem is proposed to implement three different multiobjective strategies simultaneously to make the algorithm robust to data with various features by new chain evolution rules. A local communication membrane subsystem is also designed to increase the population diversity of cluster centers by communication rules.

(3) We propose a new dense representation multisize ensemble strategy to improve the accuracy of the final results in the global ensemble membrane subsystem.

II. PROBLEM STATEMENT

A. CLUSTER ANALYSIS

Clustering is used to divide a set of objects, where objects in the same group are more similar to each other than to objects in different groups. Fuzzy C-means (FCM) [29] is one of the most popular clustering algorithms and is based on the fuzzy set principle. FCM evolves a partition matrix $U(X)$ during computation and minimizes equation (1).

$$J_m = \sum_{j=1}^{N} \sum_{i=1}^{I} \mu_{i,j}^m d^2 \left(O_i, x_j\right) \tag{1}$$

where I is the number of clusters, N denotes the number of data, O_i is the center of the ith cluster, and m is a fuzzy factor. $u_{i,j} \in U$ is an element of the fuzzy membership matrix U, which represents the membership of data x_j in O_i. $d^2 \left(O_i, x_j\right)$ denotes the distance between x_j and O_i.

By minimizing J_m with a Lagrange multiplier algorithm [30], the updating equations of memberships and cluster centers are shown as equation (2) and (3), respectively:

$$\mu_{k,j} = \frac{\left(1/d\left(O_k, x_j\right)\right)^{1/(m-1)}}{\sum_{i=1}^{I} \left(1/d\left(O_i, x_j\right)\right)^{1/(m-1)}} \tag{2}$$

$$O_k = \frac{\sum_{j=1}^{N} \mu_{k,j}^m x_j}{\sum_{j=1}^{N} \mu_{k,j}^m} \tag{3}$$

To increase the performance of FCM, multiobjective optimization has been adopted [5], [6]. In general, the Pareto solution has been used to evaluate the final results. A general multiobjective optimization problem can be expressed as follows:

$$F\left(P\right) = \left(F_1\left(P\right), F_2\left(P\right), \ldots, F_M\left(P\right)\right)^T \tag{4}$$

where F is an objective vector and M is the number of objectives. $P = (P_1, P_2, \ldots, P_D)$ is a decision vector with D-dimension.

In a Pareto relationship, solutions are divided into nondominated solutions and dominated solutions. All nondominated solutions can be considered acceptable [31].

To improve the accuracy of clustering and adapt the algorithm to data with various features, ensemble clustering uses integrated learning techniques to obtain a new clustering result by learning multiple clustering results of the merged dataset. First, the ensemble clustering implements several same/different clustering algorithms to form a base clustering pool. Then, a consistency function is employed to integrate the cluster members to obtain a unified clustering result.

B. CHAIN AND HYPERGRAPH STRUCTURES

A q-simplex (cell) s^q is the convex hull of $q+1$ affinely independent points denoted by $a_0, a_1, \ldots, a_q$. The integer q is called dimension of simplex s^q , while $a_0, a_1, \ldots, a_q$ are called vertices. A simplex s^q is uniquely indicated by its vertices and can be expressed by $[a_0, a_1, \ldots, a_q]$.

A simplicial complex P is a collection of non-empty simplices $\sigma_1, \sigma_2, \ldots, \sigma_P$, where $\sigma_1 \prec \sigma_2$ denotes that simplex σ_1 is a vertex or a face of simplex σ_2. Each simplex is supposed to be oriented [24]. Therefore, a P-chain is a kind of simplicial complex with p-dimensional simplices, defined as follows:

$$C(P) = \sum_{i=1}^{P} c_{\sigma_i}\sigma_i \tag{5}$$

where c_{σ_i} indicates the direction of simplex σ_i in chain $C^{(P)}$. Here, $c_{\sigma_i} = 1$ and $c_{\sigma_i} = -1$ represent the two directions of chain $C^{(P)}$. All simplexes of the same dimension can form a chain domain.

A hypergraph $H = (v, e)$ is a generalization of a graph whose edges contain an arbitrary number of vertices [32]–[34]. v is a set of vertices, and e is a set of hyperedges. A hyperedge can contain more than two vertices and can be formally represented by a nonempty subset of v. As shown in Fig. 1, hyperedge e_3 contains vertices v_3, v_5 and v_6.

A hypergraph $H = (v, e)$can also be described as an accessible matrix:

$$H = \begin{cases} 1, & v \in e \\ 0, & otherwise \end{cases} \tag{6}$$

where $H = 1$, if hyperedge e contains vertex v. For example, Fig.1 can be represented by:

$$\begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \\ v_7 \end{array} \begin{bmatrix} e_1 & e_2 & e_3 & e_4 \\ 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

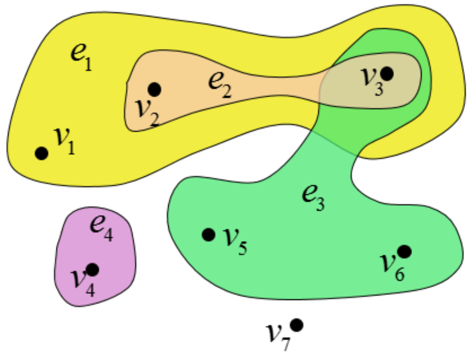


FIGURE 1. An Example of a hypergraph with vertices v={v1, v2, v3, v4, v5, v6, v7}, and hyperedges e={e1, e2, e3, e4}={{v1, v2, v3}, {v2, v3}, {v3, v4, v5}, {v4}}.

C. TISSUE-LIKE P SYSTEMS

A tissue-like P system [35], [36] associates a graph structure consisting of nodes corresponding to cells and the environment and edges that represent channels linking various components. A tissue-like P system of degree $m > 0$ with symport/antiport rules is formally defined as a tuple:

$$\prod = (O, w_1, \ldots, w_q, R_1, \ldots, R_q, i_0)\,, \tag{7}$$

where O is a finite set of objects; $w_1, \ldots, w_q$ are initial multisets of objects; R_i are finite sets of evolution rules in cell, with $1 \leq i \leq q$; and $i_0 \in \{0, 1, \ldots, q\}$ indicates the output cells of the system. R_i contains symport and antiport rules. A symport rule has the form $(i, u/\lambda, j)$, which means that the multiset of objects u goes from cell i to cell j. An antiport rule has the form $(i, u/v, j)$, indicating that the multiset of objects u in cell i and the multiset of objects v in cell j are interchanged.

The m cells are computing units that work in parallel. The tissue-like P system starts with the initial multisets $w_1, \ldots, w_q$.Then, in each step, the symport or antiport rule is applied. This process is repeated until a termination condition is satisfied. When the process terminates, the final result is embodied by the output cells.

III. HYBRID CHAIN-HYPERGRAPH P SYSTEMS FOR MULTIOBJECTIVE ENSEMBLE CLUSTERING

In this section, we propose the framework of the HCHPS. Firstly, chain and hypergraph topology structures are employed; then, we further define new rules and a configuration for the HCHPS. An HCHPS of degree $m > 0$ is a construct of the form:

$$\prod = (O, \omega_1, \omega_2, \ldots, \omega_m, subsys_i, i_0) \tag{8}$$

where O is the alphabet of objects; $\omega_1, \omega_2, \ldots, \omega_m$ are strings over O, representing the multisets of objects placed in the m cells of the system at the beginning of the computation; HCHPS contains two types of cells, i.e., chain membranes and hypermembranes. $m_{h1}, \ldots, m_{h_{m1}}$ are h_{m1} numbers of hypermembranes; and $m_{c1}, \ldots, m_{c_{m2}}$ are chain membranes with number c_{m2}, $h_{m1} + c_{m2} = m$. i_0 specifies the output membrane of $\prod$. $subsys_i$ indicates the ith subsystem of the HCHPS, of the form $subsys_i = (W_i, R_i)$, where, W_i are the initial strings contained in the ith subsystem; R_i is a finite set of rules of the ith subsystem, which are defined in the illustration of the three subsystems (section III, subsection B, (1)(2)(3)).

To better describe the tasks and relationships of different membranes in HCHPS, we divided HCHPS into different subsystems. On the one hand, each subsystem consists of different membranes, rules and channels to implement a part of the multiobjective ensemble clustering independently. On the other hand, subsystems communicate with each other to obtain the final results.

A. MEMBRANE STRUCTURES OF THE HCHPS

Chain membranes and hypermembranes are the two main membrane structures of an HCPHS.

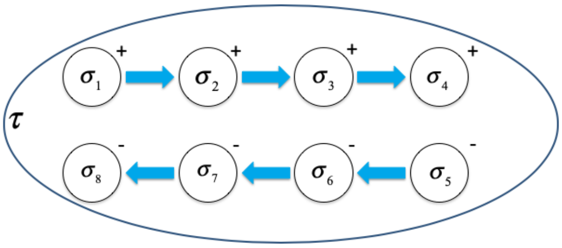


FIGURE 2. Chain membrane structure.

Definition 1: Based on the relevance operations of chain structure, membrane units combine into chains. A chain membrane is an ordered chain of connected membranes with two directions (i.e., '+' and '−'). As shown in Fig.2, $+ : \sigma_1 \rightarrow \sigma_2 \rightarrow \sigma_3 \rightarrow \sigma_4$ and $- : \sigma_5 \rightarrow \sigma_6 \rightarrow \sigma_7 \rightarrow \sigma_8$ are two chain membranes with different directions. Membranes $\sigma_1, \sigma_2, \ldots, \sigma_8$ are called unit membranes. There is a channel between any adjacent unit membranes. The outermost membrane τ is called the max membrane. Membranes $\sigma_1, \sigma_2, \ldots, \sigma_8$ including in membrane τ are called children membranes of τ. τ is the parent of $\sigma_1, \sigma_2, \ldots, \sigma_8$. $\sigma_1, \sigma_2, \ldots, \sigma_8$ and τ meets $\sigma_1, \sigma_2, \ldots, \sigma_8 \prec \tau$. A membrane without any other membranes inside it is an elementary membrane.

Definition 2: Based on the topology of hyper graph, hypermembrane is defined as a membrane with two or more upper membranes. For two membranes m_1 and m_2, m_1 is the upper membrane of m_2 if $m_2 \subset m_1$ and there is no m_3 such that $m_2 \subset m_3 \subset m_1$. m_2 is correspondingly the lower membrane of m_1. A membrane without any upper membranes is a skin membrane. A membrane without any others inside it is an elementary membrane. As shown in Fig.3, membranes 4, 5, 6, 8, 9, 10, and 11 are elementary membranes. In particular, membrane 9 is also a hypermembrane, which has upper membranes 3 and 7. Membrane 1 is the skin membrane.

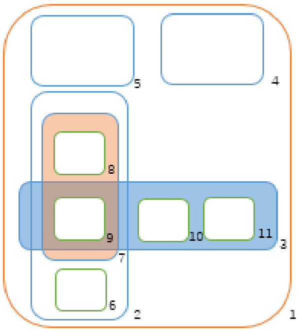


FIGURE 3. Hypermembrane structure. Membrane 1 is the skin membrane. Membranes 4, 5, 6, 8, 9, 10, and 11 are elementary membranes. Membrane 9 is a hypermembrane, which has upper membranes 3 and 7.

B. SUBSYSTEMS OF THE HCHPS

To implement multiobjective ensemble clustering in the HCHPS (HCHPS-MOEC), we propose three types of subsystems (i.e., reaction chain membrane subsystems, local communication membrane subsystems and global ensemble membrane subsystems). As shown in Fig. 4, three upper membranes implement three multiobjective clustering algorithms (i.e., NSGA-II-FCM [37], NNIA-FCM [38] and PESA-FCM [39]) simultaneously. NSGA-II-FCM, NNIA-FCM and PESA-FCM are three classic multiobjective clustering algorithms. NSGA-II-FCM is one of the best multiobjective clustering algorithms [37], which improves the overall evolution level and evolution efficiency of the middle population. However, NSGA-II-FCM is hard to maintain the diversity of the population. Different from NSGA-II-FCM, NNIA-FCM is more conducive to maintaining diversity and has higher search efficiency because of its operation of obtaining superior individuals [38]. PESA-FCM can avoid the adjustments of difficult parameters and can easily use functions with various ranges, which are not available in other multiobjective clustering algorithms [39]. To better use the advantages and remedy the disadvantages of multiobjective clustering algorithms, we selected NSGA-II-FCM, NNIA-FCM and PESA-FCM as the three multiobjective clustering algorithms in our study.

Each upper membrane has a group of local communication subsystems with a reaction chain membrane subsystem inside.

Initially, objects are generated randomly in reaction chain membrane subsystems. Each object represents a set of cluster centers. Suppose the data set to be clustered is $X = \{X_1, X_2, \ldots, X_n\} \subseteq R^{n \times d}$ with K clusters, with n being the number of data points and d the dimension. K out of n data are randomly selected as the initial cluster centers,denoted by $Q = (z_1, z_2, \ldots, z_K)$, where $Z_i = \{z_{i1}, z_{i2}, \ldots, z_{id}\} \subseteq R^d$, with $i = 1, 2, \ldots, K$. $Q = (z_1, z_2, \ldots, z_K)$ is the initial strings in the reaction chain membrane subsystem. Different chain membranes conduct multiobjective clustering algorithms with different parameters in the reaction chain membrane subsystem in parallel to obtain local nondominated objects.

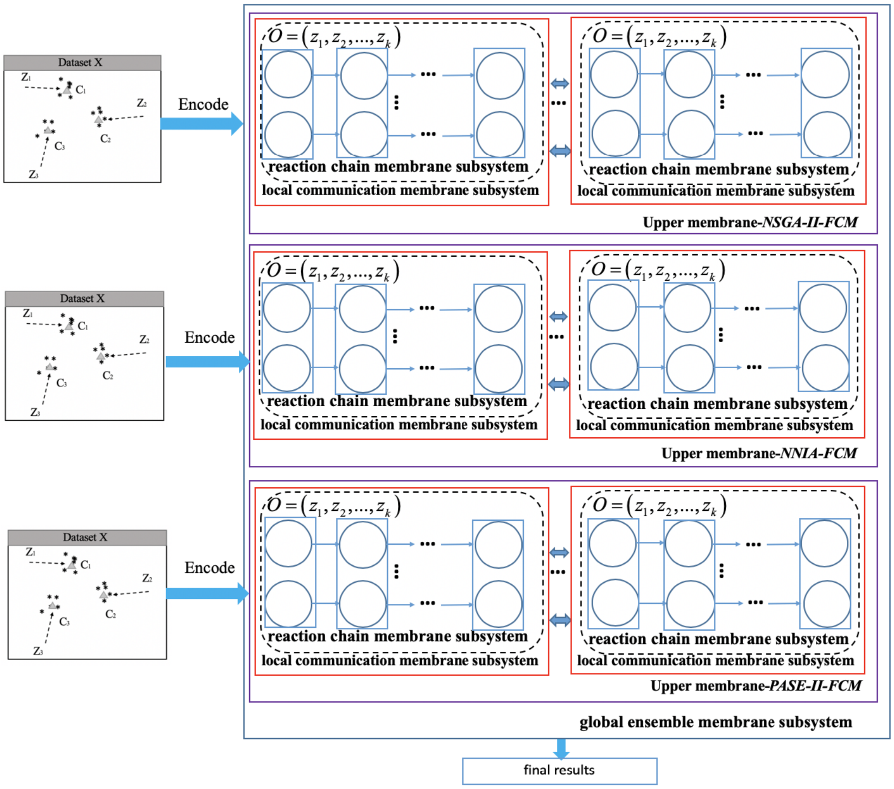

Dataset X
Z₁
C₁
Z₂
C₂
C₃
Z₃
Encode
O=(z₁,z₂,...,z_k)
reaction chain membrane subsystem
local communication membrane subsystem
Upper membrane-NSGA-II-FCM
Upper membrane-NNIA-FCM
Upper membrane-PASE-II-FCM
global ensemble membrane subsystem
final results

FIGURE 4. The membrane structures of HCHPS.

Then, the local nondominated objects are sent to the corresponding local communication subsystem to create non-dominated solutions. A semi-supervised method is adopted to select the non-dominated solutions. 20% data was given true labels before computing. Then, several optimal solutions in the Pareto set were selected based on the FM values of the 20% data and sent out from the three upper membranes to their hypermembrane. Afterwards, the final optimal solution will be obtained by the ensemble clustering in the hypermembrane, which will conduct on these optimal solutions and the remained data. To further improve the accuracy of the final results, the global ensemble membrane subsystems conduct a new dense representation multisize ensemble strategy, different numbers of optimal solutions are adopted from the Pareto set to do the ensemble clustering. The final optimal solution will be achieved from several results of the ensemble clustering.

1) REACTION CHAIN MEMBRANE SUBSYSTEM

The membrane structure of the reaction chain membrane subsystem is provided in Fig. 4. Every reaction chain membrane subsystem contains two types of membranes, i.e., relative positioning membranes and chain membranes (Fig. 2). Each chain membrane contains a set of cluster centers $Q = (z_1, z_2, \ldots, z_K)$. K is the number of clusters to be divided. The number of chain membranes is the size of the population to be evolved. Therefore, evolution of multiobjective clustering strategies can be conducted between any two chain membranes to increase the diversity of the population. Three new types of rules are designed to implement multiobjective clustering strategies in the reaction chain membrane subsystem:

- Nondominated Object Selection Rules

Nondominated object selection rules (equation (9)) are designed to select better objects from the same Pareto front by comparing the crowded distances.

$$\begin{cases} (a_k, a_l) \to a_k, & a_k^d \le a_l^d \\ (a_k, a_l) \to a_l, & a_k^d > a_l^d \end{cases} \tag{9}$$

where a_k and a_l are two objects, with $a_k \neq a_l$. a_k^d and a_l^d denote the crowded distances of a_k and a_l, respectively. Object with smaller crowded distance are selected.

- Crossover Rules

Crossover rules are designed to produce new objects for better performance, defined as:

$$\left(a_{k(i),s} + a_{k(i+1),s}\right)/2 + f \times \left(a_{k(i),s} + a_{k(i+1),s}\right) \to a_{k(i),c} \tag{10}$$

$$\left(a_{k(i),s} + a_{k(i+1),s}\right)/2 - f \times \left(a_{k(i),s} + a_{k(i+1),s}\right) \to a_{k(i),c} \tag{11}$$

where $a_{k(i),s}$ and $a_{k(i+1),s}$ are objects generated from the ith and $(i+1)$th selection, respectively, and $a_{k(i),c}$ is the object created by the ith crossover. f is a scaling factor defined in (12) $b_\mu \in [0, 1]$ and $b_\beta = 20$.

$$f = \begin{cases} 2b_\mu^{1/(b_\beta+1)}, & b_\mu \leq 0.5 \\ (2-2b_\mu)^{-1/(b_\beta+1)}, & b_\mu > 0.5 \end{cases} \tag{12}$$

- Mutation Rules

After executing the selection and crossover rules, mutation rules are conducted to increase the diversity of objects, defined as follows:

$$\begin{cases} a_k + (a_{\max} - a_{\min}) \times (2b_\mu + (1-2b_\mu) \\ \quad \times (1-(a_k - a_{\min})/(a_{\max} - a_{\min})))^{1/(b_m+1)-1} \\ \quad \to a_{k,m}, a_k \leq b_\lambda, b_\mu \leq 0.5 \\ a_k + (a_{\max} - a_{\min}) \times (2b_\mu - 1 + (2b_\mu - 1) \\ \quad \times (1-(a_{\max} - a_k)/(a_{\max} - a_{\min})))^{1/(b_m+1)} \\ \quad \to a_{k,m}, a_k > b_\lambda, b_\mu > 0.5 \end{cases} \tag{13}$$

where a_k is an object and $a_{k,m}$ is the object produced by mutation. $\alpha_{\max}$ and $\alpha_{\min}$ are the maximum and minimum values of the objects, respectively. $b_\lambda = 1/D$ is the mutation probability, D is the dimension of data set. $b_m = 20$ is the mutation parameter.

2) LOCAL COMMUNICATION MEMBRANE SUBSYSTEM

As shown in Fig. 4, local nondominated objects in different local communication membrane subsystems in the same upper membrane communicate with each other to increase the population diversity through communication rules. Then, objects with good quality are selected and sent to the corresponding upper membrane. The communication rule of the local communication membrane subsystem is:

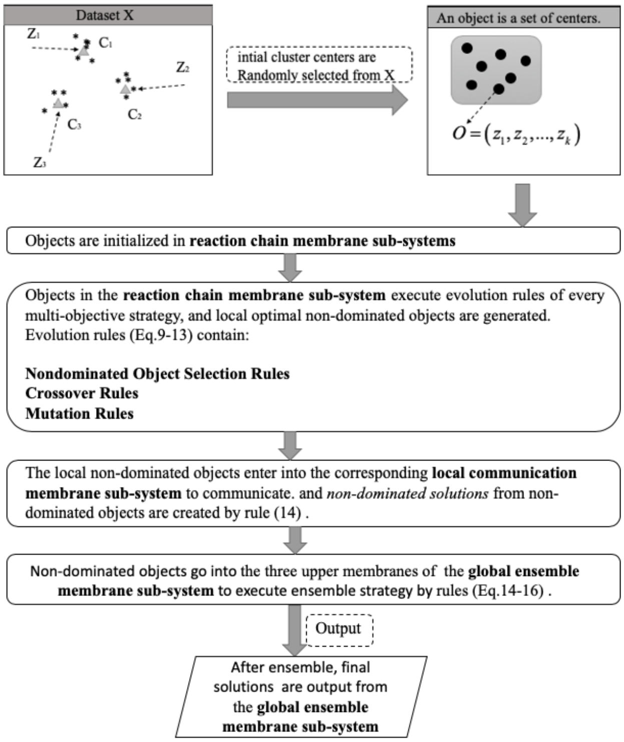
$$(p, a_k/q, a_l) \tag{14}$$

where a_k and a_l are two objects, with $a_k \neq a_l$. a_k in the local communication membrane subsystem p and a_l in q are interchanged.

3) GLOBAL ENSEMBLE MEMBRANE SUBSYSTEM

The global ensemble membrane subsystem is used to receive the global nondominated solution objects and integrate them to obtain the final clustering results. As shown in Fig. 4, the hypermembrane has three upper membranes. Each upper membrane contains a set of global nondominated solutions from three multiobjective clustering algorithms (i.e., NSGA-II-FCM, NNIA-FCM and PESA-FCM). Then, three base clustering pools are created in their upper membrane and sent to the according hypermembrane. To implement ensemble clustering, the dense representation ensemble strategy is conducted in the hypermembrane. Data objects $X = \{X_1, X_2, \ldots, X_n\}$in base clusters $Q = \{Q_1, Q_2, \ldots, Q_q\}$ are integrated into microcluster objects by rule (14). q is the number of base clusters in hypermembrane. Subsequently, consistency functions are calculated through rules (15) and (16) to produce the final results.

$$\left(p_{up}, (X_i, Q_h)/p_{hy}, (X_i, Q_h)\right) \tag{15}$$



Dataset X
Z₁
C₁
Z₂
C₂
C₃
Z₃
intial cluster centers are Randomly selected from X
An object is a set of centers.
$O=(z_1,z_2,\ldots,z_k)$
Output

Objects are initialized in **reaction chain membrane sub-systems**

Objects in the **reaction chain membrane sub-system** execute evolution rules of every multi-objective strategy, and local optimal non-dominated objects are generated. Evolution rules (Eq.9-13) contain:

Nondominated Object Selection Rules
Crossover Rules
Mutation Rules

The local non-dominated objects enter into the corresponding **local communication membrane sub-system** to communicate. and *non-dominated solutions* from non-dominated objects are created by rule (14) .

Non-dominated objects go into the three upper membranes of the **global ensemble membrane sub-system** to execute ensemble strategy by rules (Eq.14-16) .

After ensemble, final solutions are output from the **global ensemble membrane sub-system**

FIGURE 5. The flowchart of the multiobjective ensemble clustering.

The data value X_i of base cluster Q_h is sent into hypermembrane p_{hy} from upper membrane p_{up}; here, $1 \leq i \leq n$, and $1 \leq h \leq q$.

$$\left(X_i, X_j\right) \to \left(X_{i,y}, X_{j.y}\right) \tag{16}$$

Two data values X_i and X_j, $1 \leq i \neq j \leq n$, are in the same microcluster y, denoted as $X_{i,y}$ and $X_{j.y}$, if and only if they appear in the same cluster for all the base clusters.

$$y \to y_f, (PTA, PTGP) \tag{17}$$

Microcluster y changes into the final cluster y_f through two consistency functions (i.e., probability trajectory accumulation (PTA) and probability trajectory based graph partitioning (PTGP)) [10]. Specifically, in PTA, the probability trajectory based similarity of the microclusters is calculated firstly. Then, microclusters are merged to form several new clusters. This process will continue until the similarity being the maximum. Final results will be obtained by mapping microclusters back to data. In PTGP, the similarity of the microclusters is computed to construct the microcluster bipartite graph. Afterwards, the microcluster bipartite graph is partitioned into K clusters. Final results will be obtained by mapping K clusters back to data.

To further improve the clustering performance, we also integrate the ensemble results under different ensemble sizes (i.e., multisize) in the global ensemble membrane subsystem in parallel. The ensemble size is the number of solutions taking part in the ensemble learning.

TABLE 1. Seventeen real-life data sets from UCI.

No.	Name	Number of instances	Number of classes (k)	Number of dimensions (d)	Size of the classes
1	Iris	150	3	4	50,50,50
2	Wine	178	3	13	59,71,48
3	Newthyroid	215	3	5	150,35,30
4	Diabetes	768	2	8	268,500
5	Bupa	345	2	6	145,200
6	Yeast	1484	10	8	463,429,244,163,51, 44,37,30,20,5
7	Glass	214	6	9	29,76,70,17,13,9
8	Cancer	683	2	9	444,239
9	Heartstatlog	270	2	13	150,120
10	Balancescale	625	3	4	492,288,288
11	Seeds	210	3	7	70,70,70
12	Aggregation	788	6	2	170,34,273,102,130,45
13	Vowel	871	6	3	72,89,172,151,207,180
14	WBC	683	2	9	444,239
15	Ecoli	336	8	8	143,77,2,2,259,20,5,52
16	Zoo	101	7	17	41,20,5,13,4,8,10
17	heart	920	4	14	303,294,23,200

C. TERMINATION AND OUTPUT

The computation processes described above are repeated iteratively until a prescribed maximum number of computation iterations is reached. Then, the HCHPS halts. When the system halts, all the objects in the output membrane are regarded as the final HCHPS solution.

IV. EXPERIMENTS AND DISCUSSION

A. DATA SETS

We compared the proposed algorithm with state-of-the-art methods on 17 real-life data sets. Two artificial data sets are also created to visualize the clustering results. Data sets with different characteristics of shape, size, compactness and symmetry are described as follows.

1) ARTIFICIAL DATA SETS

A1 has a spiral distribution with two clusters as shown in Fig. 6(a). A2 has five clusters as shown in Fig. 6(b).

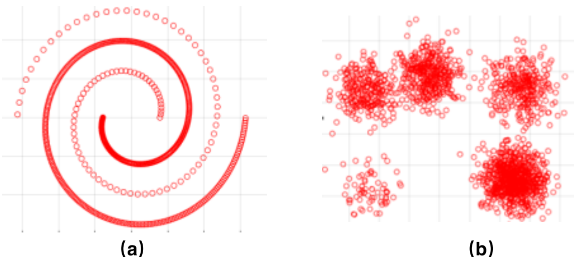

(a)
(b)

FIGURE 6. Artificial data sets: (a) A1, (b) A2.

2) REAL-LIFE DATA SETS

Seventeen real-life datasets (Table 1) from the UCI database for machine learning [28] are employed in the experiments.

B. EVALUATION METRIC

To evaluate the clustering performance, the $F - measure$ of cluster k with respect to class l is calculated by

$$F - measure\,(k, l) = \frac{2 \times (P\,(k, l) \times R\,(k, l))}{(P\,(k, l) + R\,(k, l))} \tag{18}$$

where $P\,(k, l) = s_{kl}/s_k$ and $R\,(k, l) = s_{kl}/s_l$. $P\,(k, l)$ denotes the precision of cluster k with respect to class l, and $R\,(k, l)$ denotes the recall of cluster k with respect to class l. s_{kl} is the number of data points belonging to both cluster k and l, s_k is the number of data points in k, and s_l is the number of data points in k. Therefore, $F - measure$ for the whole partitioning is calculated as

$$F - measure = \sum_{l} \frac{s_l}{s} \max F - measure\,(k, l) \tag{19}$$

where $0 \leq F - measure \leq 1$; $F - measure = 0$ indicates that all data are grouped in the wrong clusters, while $F - measure = 1$ implies exactly correct clustering.

C. PARAMETER SETTINGS

Experiments are implemented in MATLAB 2016a (MathWorks, Natick, MA). The iteration number is set as 30. The parameters are set as in Table 2. The average performance is adopted as the final performance.

D. EVALUATIONS ON DIFFERENT DATA SETS

1) COMPARISON WITH THE STATE-OF-THE-ART METHODS ON REAL-LIFE DATA SETS

In this subsection, we compare our proposed method with four state-of-the-art methods on 8 UCI data sets, as briefly introduced below.

TABLE 2. Parameter settings of HCHPS-MOEC. (*D* is the dimension of the data set.)

Upper membranes	Parameters	Description	Setting
upper membrane-NSGA-II-FCM, upper membrane- PASE-II-FCM	Pop_size	Population size	200
	Max_g	Maximum number of generations	100
	b_{α}	Crossover probability	1
	b_{λ}	Mutation probability	$1/D$
	b_{β}	Crossover parameter	20
	b_{m}	Mutation parameter	20
upper membrane- NNIA-FCM,	Pop_size	Population size	200
	c_{β}	Differential evolution parameter	0.5
	c_{m}	Mutation parameter	20

TABLE 3. Comparison of the proposed method with four state-of-the-art methods for multiobjective clustering on 8 UCI data sets. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	HCHPS-MOEC	MOFC-TMS [16]	GenclustMOO [7]	MOmoDEFC [6]	VAMOSA [5]
Iris	**1.00 ± 0.000**	0.84 ± 0.013	0.82 ± 0.028	0.81 ± 0.021	0.68 ± 0.032
Newthyroid	**0.92 ± 0.011**	0.85 ± 0.012	0.85 ± 0.025	0.84 ± 0.022	0.78 ± 0.033
Wine	**0.95 ± 0.009**	0.70 ± 0.013	0.70 ± 0.029	0.69 ± 0.025	0.52 ± 0.032
Diabetes	**0.79 ± 0.014**	0.71 ± 0.013	0.70 ± 0.024	0.68 ± 0.023	0.32 ± 0.032
Bupa	**0.78 ± 0.012**	0.68 ± 0.013	0.68 ± 0.025	0.66 ± 0.021	0.46 ± 0.034
Yeast	**0.60 ± 0.018**	0.59 ± 0.011	0.58 ± 0.027	0.54 ± 0.022	0.45 ± 0.031
Glass	**0.54 ± 0.011**	0.50 ± 0.013	0.50 ± 0.029	0.47 ± 0.023	0.42 ± 0.032
Cancer	**0.97 ± 0.010**	0.97± 0.014	0.97 ± 0.013	0.94 ± 0.023	0.82 ± 0.035

TABLE 4. Comparison of the proposed method with single-objective clustering algorithms on 8 UCI data sets. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	HCHPS-MOEC	DE [30]	PSO [31]	FCM [32]
Iris	**1.00 ± 0.000**	0.77 ± 0.021	0.76 ± 0.023	0.72 ± 0.033
Newthyroid	**0.92 ± 0.011**	0.76 ± 0.021	0.73 ± 0.023	0.66 ± 0.032
Wine	**0.95 ± 0.009**	0.65 ± 0.022	0.52 ± 0.021	0.45 ± 0.033
Diabetes	**0.79 ± 0.014**	0.77 ± 0.022	0.72 ± 0.031	0.69 ± 0.041
Bupa	**0.78 ± 0.012**	0.61 ± 0.019	0.60 ± 0.023	0.56 ± 0.032
Yeast	**0.60 ± 0.018**	0.56 ± 0.023	0.48 ± 0.025	0.44 ± 0.035
Glass	**0.54 ± 0.011**	0.46 ± 0.022	0.45 ± 0.024	0.42 ± 0.033
Cancer	**0.97 ± 0.010**	0.86 ± 0.023	0.83 ± 0.022	0.74 ± 0.034

MOFC-TMS [16]: a tissue-like P-system-based multiobjective fuzzy clustering algorithm to optimize three objectives simultaneously, which achieved the state-of-the-art performance in terms of $F - measure$.

VAMOSA [5]: a multiobjective clustering method with two objective functions being optimized together.

GenClustMOO [7]: a multiobjective clustering method that adopts a simulated annealing-based algorithm as the underlying optimization strategy.

MOmoDEFC [6]: a multiobjective clustering technique with *XB*-index and J_m being optimized together.

Table 3 compares the clustering performance of the proposed method with four state-of-the-art methods, using $F - measure$. The mean $F - measure$ increases from 0.84 to 1.00 on Iris, 0.85 to 0.92 on Newthyroid, 0.87 to 0.95 on Wine, 0.71 to 0.79 on Diabetes, 0.68 to 0.78 on Bupa, 0.59 to 0.60 on Yeast, 0.50 to 0.54 on Glass. For Cancer, the standard deviation also decreases from 0.014 to 0.010.

As can be observed in Table 3, the clustering results of HCHPS-MOEC are higher by the proposed method despite the diverse shape and size of datasets.

2) EVALUATION ON TWO ARTIFICIAL DATA SETS

As shown in Fig. 7(a) and Fig. 7(b), HCHPS-MOEC has optimal results on $A1$ and $A2$. The mean $F - measures$ are

TABLE 5. Comparison of the proposed method with our two single-objective versions on seventeen UCI data sets. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	HCHPS-MOEC	HCHPS (J_m)	HCHPS(XB)
Iris	**1.00 ± 0.000**	0.82 ± 0.016	0.82 ± 0.017
Newthyroid	**0.92 ± 0.011**	0.85 ± 0.018	0.85 ± 0.018
Wine	**0.95 ± 0.009**	0.77 ± 0.018	0.79 ± 0.019
Diabetes	**0.79 ± 0.014**	0.78 ± 0.016	0.78 ± 0.014
Bupa	**0.78 ± 0.012**	0.66 ± 0.015	0.68 ± 0.015
Yeast	**0.60 ± 0.018**	0.59 ± 0.020	0.57 ± 0.020
Glass	**0.54 ± 0.011**	0.49 ± 0.015	0.42 ± 0.016
Cancer	**0.97 ± 0.010**	0.94 ± 0.014	0.96 ± 0.014
Heartstatlog	**0.84 ± 0.018**	0.81 ± 0.019	0.81 ± 0.018
Balancescale	**0.90 ± 0.016**	0.82 ± 0.018	0.82 ± 0.019
Seeds	**1.00 ± 0.010**	0.90 ± 0.014	0.90 ± 0.013
Aggregation	**0.97 ± 0.014**	0.86 ± 0.017	0.90 ± 0.016
Vowel	**0.93 ± 0.014**	0.87 ± 0.016	0.87 ± 0.016
WBC	**0.96 ± 0.009**	0.93 ± 0.014	0.95 ± 0.013
Ecoli	**0.91 ± 0.012**	0.84 ± 0.016	0.85 ± 0.019
Zoo	**0.88 ± 0.016**	0.83 ± 0.019	0.82 ± 0.020
Heart	**0.83 ± 0.014**	0.79 ± 0.018	0.80 ± 0.017

TABLE 6. Comparison of the proposed method with three multiobjective clustering algorithm used in HCHPS on 17 UCI data sets. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	HCHPS-MOEC	NSGA-II-FCM [36]	NNIA-FCM [37]	PESA-FCM [38]
Iris	**1.00 ± 0.000**	0.98 ± 0.014	0.94 ± 0.017	1 ± 0.014
Newthyroid	**0.92 ± 0.011**	0.90 ± 0.015	0.88 ± 0.018	0.91 ± 0.015
Wine	**0.95 ± 0.009**	0.86 ± 0.018	0.77 ± 0.020	0.89 ± 0.014
Diabetes	**0.79 ± 0.014**	0.77 ± 0.019	0.73 ± 0.021	0.74 ± 0.018
Bupa	**0.78 ± 0.012**	0.65 ±0.018	0.65 ± 0.018	0.62 ± 0.016
Yeast	**0.60 ± 0.018**	0.54 ± 0.019	0.49 ± 0.020	0.57 ± 0.022
Glass	**0.54 ± 0.011**	0.43 ± 0.012	0.52 ± 0.014	0.54 ±0.012
Cancer	**0.97 ± 0.010**	0.96 ± 0.017	0.95 ± 0.019	0.96 ± 0.015
Heartstatlog	**0.84 ± 0.018**	0.82 ± 0.016	0.82 ± 0.017	0.80 ± 0.016
Balancescale	**0.90 ± 0.016**	0.89 ± 0.016	0.82 ± 0.018	0.84 ± 0.020
Seeds	**1.00 ± 0.010**	0.98 ± 0.017	0.98 ± 0.016	0.92 ± 0.018
Aggregation	**0.97 ± 0.014**	0.91 ±0.020	0.89 ± 0.023	0.93 ±0.018
Vowel	**0.93 ± 0.014**	0.91 ±0.018	0.93 ±0.016	0.93 ±0.022
WBC	**0.96 ± 0.009**	0.94 ± 0.019	0.93 ± 0.015	0.96 ± 0.021
Ecoli	**0.91 ± 0.012**	0.90 ± 0.018	0.86 ± 0.022	0.87 ± 0.015
Zoo	**0.88 ± 0.016**	0.85 ±0.019	0.87 ±0.024	0.86 ± 0.018
Heart	**0.83 ± 0.014**	0.81 ±0.021	0.80 ± 0.023	0.76 ± 0.019

both 1, indicating that the proposed method can correctly cluster $A1$ and $A2$ during all 30 iterations.

E. ABLATION STUDY

1) EVALUATION ON THE IMPACT OF MULTIOBJECTIVE OPTIMIZATION

We compared the proposed method with three classic single-objective clustering algorithms (i.e., the differential evolution algorithm (DE) [40], particle swarm optimization (PSO) [41], and FCM [29]) on 8 UCI data sets. As seen in Table 4, the $F - measure$ values of DE, PSO and FCM are all lower than that of HCHPS-MOEC. Therefore, the multiobjective optimization of HCHPS-MOEC improves the clustering accuracy significantly.

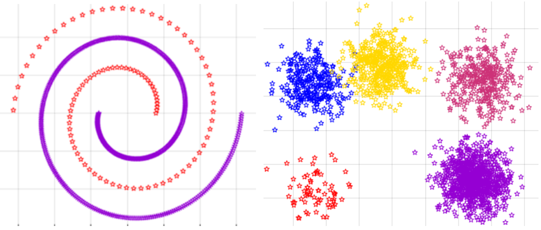
We also compared the proposed method with two single-objective algorithms (i.e., objective XB-index and J_m) implemented by the HCHPS on 17 UCI data sets. As confirmed in Table 5, multiobjective optimization is useful for clustering.

TABLE 7. Comparison of the proposed method with different ensemble sizes on 17 UCI data sets. MOCE(M)-HCHPS represents the ensemble size of the base clustering objects. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	Multisize	Size=3	Size=5	Size=7	Size=10
Iris	**1.00 ± 0.000**	0.98 ± 0.008	0.98 ± 0.0010	0.99 ± 0.014	0.98 ± 0.015
Newthyroid	**0.92 ± 0.011**	0.90 ± 0.013	0.89 ± 0.015	0.89 ± 0.017	0.89 ± 0.017
Wine	**0.95 ± 0.009**	0.88 ± 0.012	0.88 ± 0.016	0.89 ± 0.019	0.89 ± 0.021
Diabetes	**0.79 ± 0.014**	0.76 ± 0.016	0.70 ± 0.017	0.70 ± 0.019	0.70 ± 0.023
Bupa	**0.78 ± 0.012**	0.70 ± 0.015	0.68 ± 0.017	0.68 ± 0.020	0.67 ± 0.022
Yeast	**0.60 ± 0.018**	**0.60 ± 0.018**	0.58 ± 0.020	0.56 ± 0.024	0.50 ± 0.024
Glass	**0.54 ± 0.011**	0.45 ± 0.014	0.46 ±0.016	0.47 ±0.018	0.46 ±0.019
Cancer	**0.97 ± 0.010**	0.95 ± 0.012	0.95 ±0.016	0.95 ±0.019	0.95 ± 0.019
Heartstatlog	**0.84 ± 0.018**	0.82 ± 0.019	0.82 ± 0.020	0.82 ± 0.021	0.82 ±0.022
Balancescale	**0.90 ± 0.016**	0.87 ± 0.018	0.85 ± 0.019	0.85 ± 0.020	0.83 ± 0.024
Seeds	**1.00 ± 0.010**	**1.00 ± 0.010**	0.99 ±0.015	0.98 ± 0.019	0.98 ± 0.022
Aggregation	**0.97 ± 0.014**	0.90 ± 0.014	0.90 ± 0.017	0.89 ± 0.018	0.84 ± 0.019
Vowel	**0.93 ± 0.014**	0.92 ± 0.016	0.91 ± 0.018	0.91 ± 0.021	0.91 ± 0.024
WBC	**0.96 ± 0.009**	0.94 ± 0.014	0.936 ±0.017	0.93 ±0.022	0.93 ± 0.023
Ecoli	**0.91 ± 0.012**	0.90 ± 0.013	0.90 ± 0.016	0.87 ± 0.023	0.86 ± 0.025
Zoo	**0.88 ± 0.016**	0.85 ± 0.018	0.85 ± 0.023	0.83 ± 0.024	0.83 ± 0.027
Heart	**0.83 ± 0.014**	0.82 ± 0.015	0.82 ± 0.019	0.82 ± 0.022	0.82 ± 0.025

TABLE 8. Comparison of the proposed method with classic ensemble strategies on 17 UCI data sets. The best performing methods are shown in bold. (The performance is described by the mean ± std.)

UCI Datasets	HCHPS-MOEC	Voting [33]	Weighted Voting [34]	Select Voting [35]
Iris	**1.00** ± **0.000**	1.00 ± 0.018	0.97 ± 0.020	0.95 ± 0.024
Newthyroid	**0.92** ± **0.011**	0.89 ± 0.022	0.81 ± 0.028	0.79 ± 0.027
Wine	**0.95** ± **009**	0.81 ± 0.023	0.78 ± 0.026	0.77 ± 0.024
Diabetes	**0.79** ± **0.020**	0.78 ± 0.024	0.75 ± 0.028	0.70 ± 0.028
Bupa	**0.78** ± **0.012**	0.66 ± 0.025	0.74 ± 0.027	0.73 ± 0.036
Yeast	**0.60** ± **0.018**	0.55 ± 0.020	0.55 ± 0022	0.58 ± 0.022
Glass	**0.54** ± **0.011**	0.49 ± 0.022	0.41 ± 0.024	0.48 ± 0.019
Cancer	**0.97** ± **0.010**	0.96 ± 0.019	0.93 ± 0.021	0.99 ± 0.018
Heartstatlog	**0.84 ± 0.018**	0.83 ± 0.021	0.83 ± 0.022	0.77 ± 0.025
Balancescale	**0.90 ± 0.016**	0.87 ± 0.023	0.85 ± 0.020	0.82 ± 0.021
Seeds	**1.00 ± 0.010**	0.92 ± 0.019	0.90 ± 0.018	0.89 ± 0.019
Aggregation	**0.97 ± 0.014**	0.82 ± 0.020	0.86 ± 0.022	0.82 ± 0.020
Vowel	**0.93 ± 0.014**	0.83 ± 0.022	0.78 ± 0.023	0.84 ± 0.024
WBC	**0.96 ± 0.009**	0.94 ± 0.017	0.94 ± 0.018	0.82 ± 0.021
Ecoli	**0.91 ± 0.012**	0.81 ± 0.019	0.78 ± 0.020	0.70 ± 0.019
Zoo	**0.88 ± 0.016**	0.62 ± 0.021	0.71 ± 0.022	0.78 ± 0.025
Heart	**0.83 ± 0.014**	0.76 ± 0.019	0.76 ± 0.018	0.77 ± 0.018



(a) (b)

FIGURE 7. Clustering results of HCHPS-MOEC on A1 and A2.

2) EVALUATION ON THE IMPACT OF ENSEMBLE LEARNING

To verify the efficiency of the ensemble strategy, we compared the proposed method with three multiobjective clustering methods (i.e., NSGA-II-FCM, NNIA-FCM and PESA-FCM) on 17 UCI data sets. The experimental results in Table 6 show that using an ensemble strategy can improve the clustering performance.

3) EVALUATION ON DIFFERENT ENSEMBLE SIZES

Since different ensemble sizes change the fitness of data sets, thus affecting the accuracy, we compared the performance of multisize sets with fixed sizes of 3, 5, 7 and 10. Benefiting from the distribution and parallelism of P systems, the multisize experiment consumes the same time as a fixed-size experiment. As shown in Table 7, our method obtains

TABLE 9. The p-values produced by the T test compared with those of several advanced clustering approaches in terms of the F-measure.

UCI Datasets	HCHPS-MOEC vs.						
	MOFC-TMS[16]	GenclustMOO[7]	MOmoDEFC[6]	VAMOSA[5]	DE[30]	PSO[31]	FCM[32]
Iris	5.98e-167	3.22e-130	3.14e-146	3.00e-138	7.23e-151	5.24e-148	6.12e-134
	+	+	+	+	+	+	+
Newthyroid	1.50e-140	5.37e-111	1.33e-121	5.77e-115	6.16e-141	3.46e-140	1.63e-132
	+	+	+	+	+	+	+
Wine	5.77e-175	7.87e-137	1.35e-145	3.56e-132	1.25e-154	1.36e-166	4.43e-148
	+	+	+	+	+	+	+
Diabetes	3.90e-124	7.57e-115	2.79e-121	3.85e-146	1.70e-85	3.10e-98	1.44e-129
	+	+	+	+	+	+	+
Bupa	1.29e-147	2.31e-122	1.97e-133	5.30e-135	9.95e-147	4.56e-140	1.71e-128
	+	+	+	+	+	+	+
Yeast	2.67e-90	1.32e-73	1.79e-112	1.45e-119	5.07e-98	3.85e-123	1.65e-116
	+	+	+	+	+	+	+
Glass	3.39e-126	01.89e-89	3.54e-117	9.32e-115	5.79e-121	3.13e-121	3.16e-133
	+	+	+	+	+	+	+
Cancer	1	1	3.94e-94	6.82e-115	6.22e-128	1.43e-135	7.86e-127
	-	-	-	+	+	+	+

TABLE 10. Comparison of the calculation time of the proposed method with/without HCHPS on 17 UCI data sets. The best performing methods are shown in bold.

UCI Data sets	HCHPS-MOEC	MOEC
Iris	28.895 s	66.178 s
Newthyroid	23.346 s	61.322 s
Wine	21.713 s	60.584 s
Diabetes	25.357 s	70.293 s
Bupa	64.217 s	168.129 s
Yeast	67.185 s	186.003 s
Glass	25.136 s	70.182 s
Cancer	38.246 s	69.599 s
Heartstatlog	20.683 s	59.237 s
Balancescale	25.101 s	66.710 s
Seeds	19.876 s	52.856 s
Aggregation	23.831 s	64.905 s
Vowel	35.460 s	88.226 s
WBC	23.944 s	63.150 s
Ecoli	29.058 s	79.902 s
Zoo	22.335 s	58.378 s
Heart	19.720 s	52.762 s

the best results. Because the quality distribution of the base clusters is different, the number of base clusters with good quality varies. If the number of base clusters with good quality is smaller than the ensemble size, the clustering result will be disrupted by base clusters with poor quality, which will lead to poor clustering performance. Therefore, it is useful to consider the multisize ensemble for each data set.

4) EVALUATION ON DIFFERENT ENSEMBLE STRATEGIES

To verify the efficiency of different ensemble strategies, we compared the proposed method with three classic ensemble strategies, which are voting, weighted voting and select voting [42], on the 17 UCI data sets. As seen in Table 8, HCHPS-MOEC is more accurate and benefits from the finding of microclusters and use of two consensus functions.

5) SIGNIFICANCE TESTING

In this subsection, significance tests on the values of 8 UCI data sets are computed between HCHPS-MOEC and the classic clustering algorithms introduced before. The significance level is set at $p < 0.05$. The p-values are provided by Table 9, where '+' represents a significant difference and '−' represents no significant difference.

6) TIME CONSUMPTION

Table 10 provides the time consumption of multiobjective ensemble clustering with/without HCHPS systems, which shows that the parallel mode of HCHPS can improve the clustering efficiency.

V. CONCLUSION

In this paper, we have proposed a new P system with hybrid structures (HCHPS) to combine the advantages of both chain and hypergraph topology structures for multiobjective ensemble clustering. The HCHPS establishes three types of subsystems with new rules. The reaction chain membrane subsystems are used to implement three different multiobjective strategies simultaneously by new chain evolution rules. The local communication membrane subsystems are applied to increase the population diversity of cluster centers by communication rules. A new dense representation multisize ensemble strategy is conducted in the global ensemble membrane subsystem to obtain the final results. Benefiting from the parallelism of P systems, the HCHPS is less time consuming, featuring an average completion time of 28.07 seconds on 17 UCI datasets. The experimental results on 17 UCI datasets indicate that our proposed method is more accurate than previous state-of-the-art methods and remaining robust across different datasets.

REFERENCES

[1] J. A. Hartigan, *Clustering Algorithms*. Hoboken, NJ, USA: Wiley, 1975.

[2] W. Kwedlo, "Parallelizing evolutionary algorithms for clustering data," in *Proc. Int. Conf. Parallel Process. Appl. Math.* Berlin, Germany: Springer, 2005, pp. 430–438.

[3] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*, Upper Saddle River, NJ, USA: Prentice-Hall, 1988.

[4] B. S. Everitt, "Cluster analysis," *Qual. Quantity*, vol. 14, no. 1, pp. 75–100, 1980.

[5] S. Saha and S. Bandyopadhyay, "A symmetry based multiobjective clustering technique for automatic evolution of clusters," *Pattern Recognit.*, vol. 43, no. 3, pp. 738–751, 2010.

[6] I. Saha, U. Maulik, and D. Plewczynski, "A new multi-objective technique for differential fuzzy clustering," *Appl. Soft Comput.*, vol. 11, no. 2, pp. 2765–2776, 2011.

[7] S. Saha and S. Bandyopadhyay, "A generalized automatic clustering algorithm in a multiobjective framework," *Appl. Soft Comput.*, vol. 13, no. 1, pp. 89–108, Jan. 2013.

[8] N. Iam-On, T. Boongoen, S. Garrett, and C. Price, "A link-based approach to the cluster ensemble problem," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2396–2409, Dec. 2011.

[9] T. Wang, "CA-Tree: A hierarchical structure for efficient and scalable coassociation-based cluster ensembles," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 41, no. 3, pp. 686–698, Jun. 2011.

[10] D. Huang, J.-H. Lai, and C.-D. Wang, "Robust ensemble clustering using probability trajectories," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 5, pp. 1312–1326, May 2016.

[11] G. Paŭn, "Computing with membranes," *J. Comput. Syst. Sci.*, vol. 61, no. 1, pp. 108–143, Aug. 2000.

[12] J. Xue, A. Camino, S. T. Bailey, X. Liu, D. Li, and Y. Jia, "Automatic quantification of choroidal neovascularization lesion area on OCT angiography based on density cell-like P systems with active membranes," *Biomed. Opt. Express*, vol. 9, no. 7, pp. 3208–3219, 2018.

[13] J. Xue, S. Yan, J. Qu, F. Qi, C. Qiu, H. Zhang, M. Chen, T. Liu, D. Li, and X. Liu, "Deep membrane systems for multitask segmentation in diabetic retinopathy," *Knowl.-Based Syst.*, vol. 31 Jul. 2019, Art. no. 104887. doi: 10.1016/j.knosys.2019.104887.

[14] H. Peng, "Fault diagnosis of power systems using intuitionistic fuzzy spiking neural P systems," *IEEE Trans. Smart Grid*, vol. 9, no. 5, pp. 4777–4784, Sep. 2018.

[15] Z. Gao and C. Zhang, "MCIR: A multi-modal image registration algorithm based on membrane computing," in *Proc. Int. Conf. Comput. Intell. Inf. Syst.*, Apr. 2017, pp. 263–269.

[16] H. Peng, P. Shi, J. Wang, A. Riscos-Núñez, and M. J. Pérez-Jiménez, "Multiobjective fuzzy clustering approach based on tissue-like membrane systems," *Knowl.-Based Syst.*, vol. 125, pp. 74–82, Jun. 2017.

[17] L. Qin, F. Cheng, Z. Chen, and X. Huang, "A hybrid clustering algorithm based on P systems and Immune mechanisms," *ICIC Express Lett.*, vol. 9, no. 2, pp. 485–491, Jan. 2015.

[18] H. Peng, J. Wang, P. Shi, M. J. Pérez-Jiménez, and A. Riscos-Núñez, "An extended membrane system with active membranes to solve automatic fuzzy clustering problems," *Int. J. Neural Syst.*, vol. 26, no. 3, pp. 1–17, 2016.

[19] T. Gao, X. Liu, and L. Wang, "An improved PSO-based clustering algorithm inspired by tissue-like P system," in *Proc. Int. Conf. Data Mining Big Data*, 2018, pp. 325–335.

[20] H. Peng, J. Yang, J. Wang, T. Wang, Z. Sun, X. Song, X. Luo, and X. Huang, "Spiking neural P systems with multiple channels," *Neural Netw.*, vol. 95, pp. 66–71, Nov. 2017.

[21] H. Peng, J. Wang, M. J. Pérez-Jiménez, and A. Riscos-Núñez, "Dynamic threshold neural P systems," *Knowl.-Based Syst.*, vol. 163, pp. 875–884, Jan. 2019.

[22] H. Peng and J. Wang, "Coupled neural P systems," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 6, pp. 1672–1682, Jun. 2019.

[23] T. Song, L. Pan, T. Wu, P. Zheng, M. L. D. Wong, and A. Rodríguez-Patón, "Spiking neural P systems with learning functions," *IEEE Trans. Nanobiosci.*, vol. 18, no. 2, pp. 176–190, Apr. 2019.

[24] X. Liu and A. Xue, "Communication P systems on simplicial complexes with applications in cluster analysis," *Discrete Dyn. Nature Soc.*, vol. 2012, Apr. 2012, Art. no. 415242.

[25] J. Luan and X. Liu, "Logic operation in spiking neural P system with chain structure," in *Frontier and Future Development of Information Technology in Medicine and Education*. Amsterdam The Netherlands: Springer, 2014.

[26] B. Heintz, R. Hong, S. Singh, G. Khandelwal, C. Tesdahl, and A. Chandra, "MESH: A flexible distributed hypergraph processing system," 2019, *arXiv:1904.00549*. [Online]. Available: https://arxiv.org/abs/1904.00549

[27] P. Li and O. Milenkovic, "Inhomogeneous hypergraph clustering with applications," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2308–2318.

[28] Designated National Authorities Asuncion. (2007). *UCI Machine Learning Repository*. [Online]. Available: http://www.ics.uci.edu/mlearn/MLRepository.html

[29] C. Liu, Q. Chen, Y. Chen, and J. Liu, "A fast multiobjective fuzzy clustering with multimeasures combination," *Math. Problems Eng.*, vol. 2019, Jan. 2019, Art. no. 3821025.

[30] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*. New York, NY, USA: Plenum, 1981.

[31] K. Faceli, M. C. P. de Souto, D. S. A. de Araújo, and A. C. P. L. F. de Carvalho, "Multi-objective clustering ensemble for gene expression data analysis," *Neurocomputing*, vol. 72, nos. 13–15, pp. 2763–2774, Aug. 2009.

[32] P. Corsini and V. Leoreanu, *Graphs and Hypergraphs*. Amsterdam, The Netherlands: North Holland, 1973.

[33] T. W. Ha, J. H. Seo, and M. H. Kim, "Efficient searching of subhypergraph isomorphism in hypergraph databases," in *Proc. IEEE Int. Conf. Big Data Smart Comput.*, Jan. 2018, pp. 739–742.

[34] D. Zhou, J. Huang, and B. Schölkopf, "Learning with hypergraphs: Clustering, classification, and embedding," in *Proc. Int. Conf. Neural Inf. Process. Syst* Cambridge, MA, USA: MIT Press, 2006, pp. 1601–1608.

[35] R. Freund, G. Păun, and M. J. Pérez-Jiménez, "Tissue P systems with channel states," *Theor. Comput. Sci.*, vol. 330, no. 1, pp. 101–116, Jan. 2005.

[36] B. Song, C. Zhang, and L. Pan, "Tissue-like P systems with evolutional symport/antiport rules," *Inf. Sci.*, vol. 378, pp. 177–193, Feb. 2017.

[37] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: NSGA-II," *IEEE Trans. Evol. Comput.*, vol. 6, no. 2, pp. 182–197, Apr. 2002.

[38] M. Gong, L. Jiao, H. Du, and L. Bo, "Multiobjective immune algorithm with nondominated neighbor-based selection," *Evol. Comput.*, vol. 16, no. 2, pp. 225–255, 2008.

[39] D. W. Corne, N. R. Jerram, J. D. Knowles, and M. J. Oates, "PESA-II: Region-based selection in evolutionary multiobjective optimization," in *Proc. 3rd Annu. Conf. Genetic Evol. Comput.*, Jul. 2001, pp. 283–290.

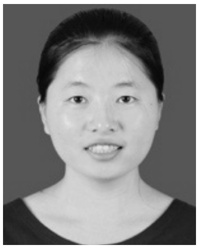
[40] K. Price, R. M. Storn, and J. A. Lampinen, *Differential Evolution: A Practical Approach to Global Optimization* (Natural Computing Series), vol. 141, no. 2. Springer, 2005.

[41] R. J. Kuo, Y. J. Syu, Z.-Y. Chen, and F. C. Tien, "Integration of particle swarm optimization and genetic algorithm for dynamic clustering," *Inf. Sci.*, vol. 195, pp. 124–140, Jul. 2012.

[42] Z.-H. Zhou and W. Tang, "Clusterer ensemble," *Knowl.-Based Syst.*, vol. 9, no. 1, pp. 77–83, Mar. 2006.

SHUO YAN received the B.S. degree in information management and information system from Shandong Normal University, in 2017, where she is currently pursuing the M.S. degree in management science and engineering.

YUAN WANG received the bachelor's degree in management from the University of Qiqihar, in 2017. She is currently pursuing the master's degree in management with Shandong Normal University, majoring in management science and engineering.

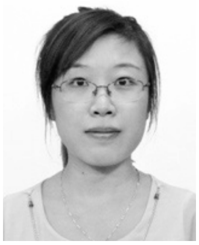


DETING KONG received the B.S. degree in information management and information system from Shandong Normal University, in 2019, where she is currently pursuing the M.S. degree in management science and engineering.

JINYAN HU is currently pursuing the B.S. degree in information management and information system with Shandong Normal University.

JIANHUA QU received the degree from Shandong Normal University, in 2000, the B.S. degree in computer science and technology, and the Ph.D. degree in information management and e-commerce from Shandong Normal University, in 2010. Since 2010, she has been an Associate Professor and the Dean of the Department of Information Management and Information System, Shandong Normal University. Her current research interests include computational intelligence and data mining.

XIYU LIU received the Ph.D. degree in mathematical sciences from Shandong University, Jinan, China, in 1990. He is currently the Dean of the Academy of Management Science and Engineering, Shandong Normal University, China. He is the author of two books and more than 140 articles. His current research interests include membrane computing, data mining, computational intelligence, and nonlinear analysis. Prof. Liu received the Taishan Scholar of Management Science and Engineering. He was the Vice President of the Computer Education Research Association of China Higher Normal Universities and the Shandong Computer Society.



JIE XUE received the B.S. and Ph.D. degrees in management science and engineering from Shandong Normal University, in 2010 and 2015, respectively. She did the National Visiting Scholar Program with the University of North Carolina, from 2017 to 2018. She is currently an Associate Professor with the Business School, Shandong Normal University. Her current research interests include machine learning, biocomputing, and medical image processing.