Evolutionary Multiobjective Clustering Algorithms with Ensemble for Patient Stratification

Yunhe Wang, Xiangtao Li, Ka-Chun Wong, Yi Chang, and Shengxiang Yang

***Abstract*—Patient stratification has been studied widely to tackle subtype diagnosis problems for effective treatment. Due to the dimensionality curse and poor interpretability of data, there is always a long-lasting challenge in constructing a stratification model with high diagnostic ability and good generalization. To address these problems, this paper proposes two novel evolutionary multiobjective clustering algorithms with ensemble (NSGA-II-ECFE and MOEA/D-ECFE) with four cluster validity indices used as the objective functions. First, an effective ensemble construction method is developed to enrich the ensemble diversity. After that, an ensemble clustering fitness evaluation (ECFE) method is proposed to evaluate the ensembles by measuring the consensus clustering under those four objective functions. To generate the consensus clustering, ECFE exploits the hybrid co-association matrix from the ensembles and then dynamically selects the suitable clustering algorithm on that matrix. Multiple experiments have been conducted to demonstrate the effectiveness of the proposed algorithm in comparison with seven clustering algorithms, twelve ensemble clustering approaches, and two multiobjective clustering algorithms on 55 synthetic datasets and 35 real patient stratification datasets. The experimental results demonstrate the competitive edges of the proposed algorithms over those compared methods. Furthermore, the proposed algorithm is applied to extend its advantages by identifying cancer subtypes from five cancer-related single-cell RNA-seq datasets.**

***Index Terms*—Multiobjective optimization, ensemble clustering, patient stratification.**

I. Introduction

PATIENT stratification is a critical task in cancer diagnosis and treatment, which aims to group patients into disease subgroups. This will lead to the development of personalised, preventive or therapeutic strategies by identifying patients who are more likely to respond positively to a given therapy. However, there still exists ground challenges in discovering cancer groupings due to sample imbalance and experimental noises [1]. Therefore, it is imperative to design efficient computational methods to stratify the patient data into cancer subtypes precisely.

In recent years, several clustering methods have been proposed to identify patient stratification data. For instance, Liu *et al.* [2] proposed a network-assisted co-clustering method to group cancer data into different subtypes. Wang *et al.* [3] introduced a consensus clustering method based on an optimization process with regularization to aggregate and differentiate patient outcomes. Ester *et al.* [4] developed an integrative Bayesian biclustering method to analyze the patient stratification datasets. Graim *et al.* [5] presented a community detection framework for choosing subtypes out of sparse patient measurements. However, only one cluster validity index is evolved in those algorithms. It is difficult for an algorithm with one internal evaluation function to be robust and interpretive for almost all datasets.

To capture different characteristics of the datasets, many multiobjective clustering approaches based on multiple cluster validity indices have been developed. Mukhopadhyay *et al.* [6] developed a novel interactive genetic algorithm-based multiobjective approach by evolving a set of clustering validity measures to cluster real-life gene expression datasets; Shi *et al.* [7] proposed a transfer clustering ensemble selection algorithm (TCES) under a multiobjective self-evolutionary process, in which three objective functions are optimized in a target dataset transferred from a source dataset; Li and Wong [8] proposed a multiobjective clustering method by fast search of density peaks (MOCDP) with five cluster validity indices served as the objective functions to stratify the patients into subtypes; Wang *et al.* [9] investigated a multiobjective spectral clustering algorithm (MOSC) for patient stratification based on decomposition under two clustering validation measures. Unfortunately, those methods always employ one clustering algorithm as the basic clustering algorithm. We hardly believe that any basic clustering algorithm can be the all-time winner for all those patient stratification data. Moreover, each clustering algorithm has its own merits and disadvantages.

Ensemble clustering techniques have attracted increasing attention and emerged as a powerful tool for patient stratification by using multiple clustering algorithms to yield better clustering performance than a single clustering algorithm. Liu *et al.* [10] developed an entropy-based consensus clustering algorithm (ECC) that merges the basic partitions into a consensus one by an entropy-based utility function. Yu *et al.* [11] proposed a projective clustering ensemble by combining the superiority of projective clustering and ensemble clustering to

Y.H. Wang is with the School of Artificial Intelligence, Jilin University, Changchun, Jilin 130012, China, with the School of Information Science and Technology, Northeast Normal University, Changchun, Jilin 130117, China, and with the School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, UK. E-mail: wangyh082@nenu.edu.cn.

X.T. Li is with the School of Artificial Intelligence, Jilin University, Changchun, Jilin 130012, China. (Corresponding author: lixt314@jlu.edu.cn.)

K.C. Wong is with the Department of Computer Science, City University of Hong Kong, Hong Kong. E-mail: kc.w@cityu.edu.hk.

Yi Chang is with the School of Artificial Intelligence, Jilin University, Changchun, Jilin 130012, China. E-mail: yichang@jlu.edu.cn.

S. Yang is with the School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, UK. E-mail: syang@dmu.ac.uk.

enhance the clustering quality. Unfortunately, most of them ignore the importance of the member diversity in an ensemble to prevent the ensemble algorithm from being trapped into a local optimum [12]. Meanwhile, in the ensemble clustering method, current co-association matrices usually focus on one modality of the ensembles, resulting in the distortion characteristic for the clustering on a single modality [13]. The data modality is defined as the data representation, which is produced by a specific process and can be used to define clusters on its own. Meanwhile, a co-association matrix can be aggregated from different clusters within the same single data modality [13]. Moreover, those methods often consider all samples to be equally important in the similarity matrix. Thus, it is quite essential to develop a co-association matrix that involves more than one data modality and weighs each sample distinctively. In this study, we propose two novel evolutionary multiobjective clustering algorithms with ensemble (NSGA-II-ECFE and MOEA/D-ECFE) to address aforementioned limitations. Firstly, an effective ensemble construction method is proposed to maintain the ensemble diversity. In order to measure the intrinsic characteristics of the ensembles, an ensemble clustering fitness evaluation (ECFE) method is developed by evaluating the consensus clustering that is generated from the ensemble. In ECFE, a hybrid co-association matrix is proposed to combine the advantages of different co-association matrices to exploit the appropriate subtype structure. In addition, we dynamically select the suitable clustering algorithm to produce the consensus clustering. To guide the evolution of those ensemble mechanisms, four cluster validity indices, including DB, Dunn, cohesion, and stability, are proposed to capture diverse characteristics of the dataset under two evolutionary multiobjective optimization techniques, namely nondominated sorting genetic algorithm II (NSGA-II) [14] and multiobjective evolutionary algorithm based on decomposition (MOEA/D) [15]. The efficiency of the proposed method is tested on 55 synthetic datasets and 35 real patient stratification datasets. The results demonstrate that our proposed algorithm significantly outperforms other approaches including seven clustering algorithms, twelve ensemble clustering methods, and two multiobjective clustering algorithms. Meanwhile, sensitive analysis and extensive experiments are performed to extend the performance of the proposed algorithm.

The main contributions of this study are summarized as follows:

- We propose an effective ensemble construction method to maintain the ensemble diversity. For each ensemble, we employ the k-means clustering method to generate half base clusterings; meanwhile we adopt the locus-based adjacency genetic scheme to produce the rest base clusterings.
- We propose an ensemble clustering fitness evaluation (ECFE) to measure the fitness of the ensemble by evaluating those four objective functions on the consensus clustering. In ECFE, to produce the consensus clustering, first, we propose and design a hybrid co-association matrix for the ensembles to represent the subtype structure of the patient stratification data. It incorporates the sample rank into two types of co-association matrix to generate a similarity matrix for the patient stratification data, which can exploit the appropriate subtype structure during the evolutionary multiobjective clustering. Then, we choose a suitable basic clustering algorithm dynamically using the multiobjective evolutionary optimization to generate the consensus clustering from the hybrid co-association matrix.
- Four cluster validity indices including DB, Dunn, cohesion, and stability are optimized simultaneously to guide the multiobjective clustering, capturing various characteristics of the evolved clusterings.
- Experiments on 55 synthetic datasets and 35 real patient stratification datasets show that the proposed algorithms significantly outperform compared existing methods.

The rest of this study is organized as follows. The problem formulation and multiobjective optimization are summarized in Section II. The proposed method is detailed in Section III. The experimental design and results are presented in Section IV. The extended application of the proposed algorithm is outlined in Section V. Finally, the conclusion and future works are provided in Section VI.

II. Preliminaries

A. Problem Formulation

Let $X = \{x_1, x_2, ..., x_i, ..., x_n\}$ be a patient stratification dataset of n data samples, where $x_i = \{x_i^1, x_i^2, ..., x_i^m\}$, $(i \in \{1, 2, ..., n\})$, m is the number of genes. The ensemble clustering problem is to build a consensus clustering $\pi_* = \left\{C_*^1, C_*^2, ..., C_*^N\right\}$ using the information of multiple base clusterings (ensemble) Π, where N is the number of clusters in the final clustering of X. The ensemble can be denoted as $\Pi = \{\pi_1, \pi_2, ..., \pi_i, ..., \pi_M\}$, where $\pi_i = \left\{C_i^1, C_i^2, ..., C_i^j, ..., C_i^{N_i}\right\}$ $(i \in \{1, 2, ..., M\})$ is the i-th base clustering, C_i^j $(j \in \{1, 2, ..., N_i\})$ is the j-th cluster in π_i, and N_i is the number of clusters in π_i. Given $j, k \in \{1, 2, ..., N_i\}, j \neq k$, it holds that $C_i^j \bigcap C_i^k = \varnothing$ since each data point can only belong to one cluster in a base clustering. Let $C = \left\{C^1, C^2, ..., C^{N_c}\right\}$ be the set of all clusters in Π, it is obvious that $N_c = \sum_{i=1}^{M} N_i$.

B. Multiobjective Optimization

The multiobjective optimization (MOO) considers optimization problems involving more than one objective function to be optimized simultaneously. It can be characterized as follows:

$$\begin{aligned} min \quad & f_i(V), \quad i = 1, 2, ..., I \\ subject\ to \quad & g_j(V) \leq 0, \quad j = 1, 2, ..., J \\ & h_k(V) = 0, \quad k = 1, 2, ..., K \end{aligned} \tag{1}$$

where $V = \{v_1, v_2, ..., v_o\}$ is a feasible solution with o decision variables, I is the number of objective functions. An MOO problem is to find a solution set that optimizes those I objective functions simultaneously, satisfying all J equality and K inequality constraints. Considering a minimization problem, V_1 is said to dominate V_2, if for all i, $f_i(V_1) \leq f_i(V_2)$

and for at least one i, $f_i(V_1) < f_i(V_2)$. V^* is termed as a Pareto-optimal (non-dominated) solution if and only if there does not exist a solution V' that dominates V^*. The set of all those non-dominated solutions is called the Pareto set (PS).

III. Proposed Method

A. *Methodology Overview of ECFE*

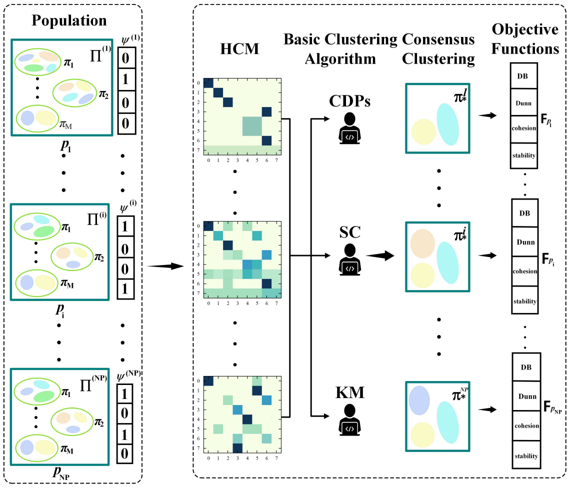


Fig. 1: An overview of ECFE.

The ECFE method is proposed to evaluate the ensembles in the proposed algorithm. Considering a patient stratification dataset X with n samples and m genes, to reduce the number of genes in X, we firstly filter out genes with low variance [16] and retain D genes in X. Then, a population $P = \{p_1, p_2, ..., p_i, ..., p_{NP}\}$ with NP individuals is constructed. We design each individual with an ensemble including a pool of M base clusterings and a random parameter vector, which is denoted as $p_i = \left\{\Pi^{(i)}, \Psi^{(i)}\right\} = \left\{\pi_1^{(i)}, \pi_2^{(i)}, ..., \pi_M^{(i)}, \alpha^{(i)}, \beta_1^{(i)}, \beta_2^{(i)}, \beta_3^{(i)}\right\}$, where $i \in \{1, 2, ..., NP\}$; $\pi_{j=1,2,...,M}^{(i)}$ is a base clustering; $\alpha^{(i)}, \beta_{k=1,2,3}^{(i)} \in \{0, 1\}$; $\sum_{k=1}^{3} \beta_k^{(i)} = 1$; $\beta_k^{(i)} = 0$ means that the k-th basic clustering algorithm in a pool of base clustering algorithms is chosen; otherwise, it represents that the k-th basic clustering algorithm is not chosen; α is the parameter to calculate the similarity between different clustering members in $\Pi^{(i)}$. An overview of ECFE is depicted in Fig. 1. From Fig. 1, the purpose of fitness evaluation is to give a quantitative indicator to determine who is eligible to be parent solutions in the multiobjective evolutionary algorithm. This fitness evaluation method is sufficient to investigate the tendency of the ensemble performance. If ensembles are with better performance, they will probably be able to generate a high-quality consensus clustering. In ECFE, the hybrid co-association matrix (HCM) is proposed to represent the subtype structure from the fuse information of each individual. Next, a basic clustering algorithm is selected dynamically using each individual from a pool of base clustering algorithms, in which three basic clustering algorithms including spectral clustering (SC) [17], k-means (KM) [18], and clustering by fast search of density peaks (CDPs) [19] are considered to produce the consensus clustering. Finally, we measure the quality of the ensemble in the population by evaluating the objective functions on the consensus clustering. It can be observed that the HCM and the basic clustering algorithm can be optimized by evolving the population in the multiobjective evolutionary optimization.

1) Construction of the Ensemble: The base clusterings generated by any single method in the ensemble are usually similar. To enhance the performance of ensemble clustering, diverse ensembles have been proven to be effective for addressing this issue [12]. In this study, we propose to construct the ensemble using two generation methods, as depicted in Fig. 2. First, we use the k-means algorithm to produce half of the base clusterings. The number of clusters is selected randomly from [2,$\sqrt{n}$] [20], where n is the number of samples in the patient stratification dataset. Then, we adopt the locus-based adjacency genetic scheme [21] for the other half base clusterings. However, there still exists some redundant clusterings in the base clusterings, resulting in a loss of diversity. Therefore, to maintain the diversity of the base clusterings, a relabel strategy is employed to recode the base clusterings. Given a base clustering $\pi = \left\{C^1, C^2, ..., C^N\right\}$ in the ensemble Π that is divided into N clusters, the relabel strategy is to align each partition of the clustering from the numbered cluster 1 to the numbered cluster n sequentially. For example, given two clusterings $\pi_1 = \{3, 3, 3, 1, 1, 2, 2\}$ and $\pi_2 = \{2, 2, 2, 1, 1, 3, 3\}$ in the ensemble $\Pi^{(1)}$, by aligning from 1 to 3 sequentially, $\pi_1 = \{1, 1, 1, 2, 2, 3, 3\}$ and $\pi_2 = \{1, 1, 1, 2, 2, 3, 3\}$ are shaped over the relabel strategy. Obviously, they are duplicate clusterings. Once we find the redundant clustering, we will assign random clusterings to them. The random clustering provides each sample with a cluster chosen randomly from 1 to N. It is worth noting that the relabel strategy is adopted to eliminate some redundant clusterings during the evolution process.

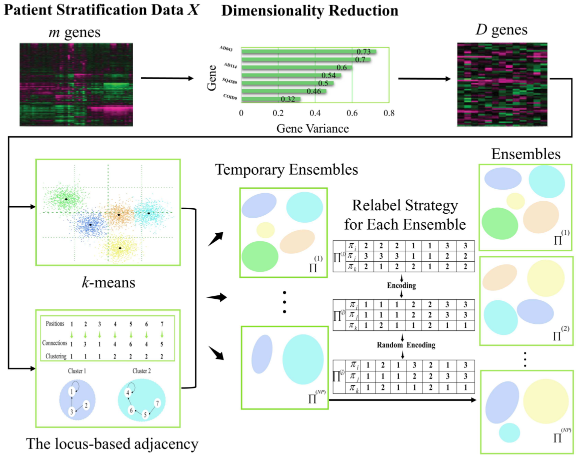


Fig. 2: Construction of the ensembles. First, dimensionality reduction is adopted based on the variances of those m genes in the patient stratification dataset X. Then, an ensemble is constructed from the remaining D genes. For each ensemble, the k-means method is used to produce half clusterings and the locus-based adjacency genetic scheme is used to produce the rest clusterings. After that, the relabel strategy is employed on each ensemble, in which random clusterings are generated to replace those redundant clusterings in that ensemble.

2) Hybrid Co-association Matrix: In [22], an evidence accumulation co-association matrix (EACM) was proposed to cope with ensemble members with different numbers of clusters, which considers a measure of similarity between base clusterings, defined as follows:

$$EACM(i,j) = \frac{1}{M}\sum_{p=1}^{M}\sum_{q=1}^{N_p}\Gamma(i,j,C_p^q) \tag{2}$$

where C_p^q is the q-th cluster in the base clustering π_p, $\Gamma(i,j,C_p^q)$ is an indicator and can be defined:

$$\Gamma(i,j,C_p^q) = \begin{cases} 1, & if\ x_i \in C_p^q \wedge x_j \in C_p^q \\ 0, & otherwise \end{cases} \tag{3}$$

However, such co-association matrix ignores the potential relationship between different clusters. An enhanced co-association matrix [23] (ECM) was developed to capture the sample-wise relationship and the cluster-wise similarity simultaneously, which is defined as:

$$ECM(i,j) = \frac{1}{M}\sum_{p=1}^{M}\sum_{q=1}^{N_p}\Gamma_E(i,j,C_p^q) \tag{4}$$

$$\Gamma_E(i,j,C_p^q) = \begin{cases} 1, & if\ x_i \in C_p^q \wedge x_j \in C_p^q \\ z_{uv}, & otherwise \end{cases} \tag{5}$$

$$\begin{aligned} z_{uv} &= \frac{R_{u:}^{(1:t)} \cdot R_{v:}^{(1:t)}}{\left\| R_{u:}^{(1:t)} \right\|_2 \times \left\| R_{v:}^{(1:t)} \right\|_2} \\ &= \frac{\left\langle R_{u:}^{(1:t)}, R_{v:}^{(1:t)} \right\rangle}{\sqrt{\left\langle R_{u:}^{(1:t)}, R_{u:}^{(1:t)} \right\rangle \times \left\langle R_{v:}^{(1:t)}, R_{v:}^{(1:t)} \right\rangle}} \\ R^{(t)} &= \left\{ r_{uv}^{(t)} \right\}_{n\times n} = \begin{cases} R, & if\ t=1 \\ R^{(t-1)}\cdot R, & if\ t>1 \end{cases} \\ r_{uv} &= \begin{cases} \frac{e_{uv}}{\sum_{C^k \neq C^u} e_{uk}}, & if\ u \neq v \\ 0, & if\ u = v \end{cases} \\ e_{uv} &= \frac{|C^u \cap C^v|}{|C^u \cup C^v|} \end{aligned} \tag{6}$$

where $x_i \in C_p^u$ and $x_j \in C_p^v$, $p \in \{1,2,....M\}$,$u,v \in \{1,2,...,N_c\}$, $z_{uv} \in [0,1]$ is the cluster-wise similarity, $\langle\cdot,\cdot\rangle$ is the dot product of two vectors, t is the number of steps of the random walks, $R_{u:}^{(1:t)} = \left\{R_{u:}^{(1)}, R_{u:}^{(2)}, ..., R_{u:}^{(t)}\right\}$ ($R_{u:}^{(t)} = \left\{r_{u1}^{(t)}, r_{u2}^{(t)}, ..., r_{un}^{(t)}\right\}$) is the random walk trajectory from step 1 to step t for the random walker that starts from C^u, $R^{(t)}$ is the multistep transition probability matrix, $\bigcap$ represents the intersection of two sets, $\bigcup$ is the union of two sets, and $|\cdot|$ is the number of elements in a set. Since the entries in $R^{(t)}$ are non-negative, z_{uv} is naturally within the range [0,1] [23]. In particular, z_{uv} is equal to 0 when $R_{u:}^{(1:t)} \cdot R_{v:}^{(1:t)} = 0$; z_{uv} is equal to 1 when $R_{u:}^{(1:t)} = R_{v:}^{(1:t)}$, which is proven in Supplementary Section II.

In fact, a single similarity matrix cannot always represent the cluster structure of all patient stratification datasets very well. Moreover, each sample is treated equally in the similarity matrix, neglecting the demand for non-isometric distances between pair-wise samples. Therefore, to convey more sample-wise information of the patient stratification data, a hybrid co-association matrix (HCM) is proposed in this study by switching EACM and ECM alternatively:

$$\begin{aligned} HCM &= diag(W) \times (\alpha EACM + (1-\alpha)ECM) \\ &= diag(W) \times \frac{\alpha}{M}\sum_{p=1}^{M}\sum_{q=1}^{N_p}\Gamma(i,j,C_p^q) \\ &+ diag(W) \times \frac{(1-\alpha)}{M}\sum_{p=1}^{M}\sum_{q=1}^{N_p}\Gamma_E(i,j,C_p^q) \\ &= diag(W) \times \frac{1}{M} \times \sum_{p=1}^{M}\sum_{q=1}^{N_p}\Gamma_H(i,j,C_p^q) \end{aligned} \tag{7}$$

$$\Gamma_H(i,j,C_p^q) = \begin{cases} 1, & if\ x_i \in C_p^q \wedge x_j \in C_p^q \\ z_{uv}, & if\ \neg(x_i \in C_p^q \wedge x_j \in C_p^q)\ and\ \alpha = 0 \\ 0, & if\ \neg(x_i \in C_p^q \wedge x_j \in C_p^q)\ and\ \alpha = 1 \end{cases} \tag{8}$$

where α is a binary number that can be obtained from the parameter vector Ψ; $x_i \in C_p^u$ and $x_j \in C_p^v$; $u,v \in \{1,2,...,N_c\}$; $z_{uv} \in [0,1]$ is the cluster-wise similarity, the boundary value conditions of it are similar to the Eq. (6); $diag(W) = diag(1,2,...,n)$ is a diagonal matrix that represents the initial ranks for the samples in the patient stratification data that can boost the sample-wise distance in the similarity matrix.

B. Objective Functions

The suitable choice of objective functions has an important role in guiding the multiobjective optimization. For a majority of multiobjective evolutionary clustering algorithms, multiple cluster validity indices have been optimized simultaneously as objective functions. In this study, four objective functions are considered to optimize the clustering problem with ensemble, enabling the proposed algorithm to capture diverse properties of the clusters in an unsupervised way. Let $\pi = \left\{C^1, C^2, ..., C^N\right\}$ be a clustering solution, the first objective is the Davies-Bouldin (DB) index [24], which can be described as:

$$f_1(\pi) = \frac{1}{K}\sum_{i=1}^{K}\max_{i\neq j}\left(\frac{\delta_i + \delta_j}{d(c^i, c^j)}\right) \tag{9}$$

where δ_i and δ_j are the intracluster distances of cluster C^i and cluster C^j, $d(c^i, c^j)$ is the Euclidean distance between the cluster centroid c^i and c^j. It measures the intracluster similarity of π. Smaller values indicate better clustering results, namely $f_1(\pi)$ should be minimized.

The second objective is the Dunn index [25]:

$$f_2(\pi) = \min_i\left\{\min_j\left(\frac{\min_{x\in C^i, y\in C^j} d(x,y)}{\max_k \max_{x,y\in C^k} d(x,y)}\right)\right\} \tag{10}$$

where $d(x,y)$ is the Euclidean distance between two data samples x and y. It discovers compact and well-separated clusters in π. Clusterings with larger values denote better clustering results. Therefore, $f_2(\pi)$ should be maximized.

Although DB and Dunn are always used to validate the clustering performance of different clustering algorithms [26], they have not taken the cluster connectivity into consideration. Therefore, to generate a high-quality clustering, we propose to employ cohesion and stability based on the cluster density to measure the quality of each clustering.

Cohesion and stability [27], which concentrate on the density-based connectivity between C^i and other clusters in π and the inner density-based connectivity of π, are served as the third and last objectives, which are respectively defined as follows:

$$f_3(\pi) = \sum_{i=1}^{K} \max_{x_{p\in C^i}, x_{q\in(\pi\setminus C^i)}} RSIM(x_p, x_q, S_X, l) \tag{11}$$

$$f_4(\pi) = \sum_{i=1}^{K} \min_{x_{p\in C^{i1}}, x_{q\in C^{i2}}} RSIM(x_p, x_q, S_{C^i}, l) \tag{12}$$

where $RSIM(x_i, x_j, S_X, l)$ is the robust minimax similarity [28], S_X indicates the similarity matrix of X, l represents the number of the nearest neighbors, C^{i1} and C^{i2} are two different clusters. A good clustering is expected to have weak cohesion and strong stability. Therefore, $f_3(\pi)$ should be minimized and $f_4(\pi)$ should be maximized to achieve a good clustering.

C. Evolutionary Multiobjective Clustering Algorithm with Ensemble

In this section, we employ the multiobjective evolutionary algorithm to optimize those four cluster validity indices simultaneously, obtaining the appropriate data modality and the suitable basic clustering algorithm in ECFE. Two effective evolutionary multiobjective optimization techniques, NSGA-II and MOEA/D, are used to evolve the population iteratively. The proposed algorithms based on ECFE (NSGA-II-ECFE and MOEA/D-ECFE) are detailed as follows:

1) NSGA-II-ECFE: NSGA-II-ECFE starts with a collection of NP individuals. Each individual contains an ensemble with M basic clusterings and a parameter vector. We propose to evaluate the fitness of each individual by ECFE. In each iteration, a mating pool is generated using the current population under the binary tournament selection. For each individual, the uniform crossover operator and the neighborhood biased mutation operator [29] are undertaken on each clustering of the ensembles in the population to discover better clusterings in the ensemble. Those four objective functions are calculated on the new individual by the proposed ECFE to guide the multiobjective evolution. After that, a new population is produced including $2NP$ individuals. Finally, NP individuals are chosen by the fast non-dominated sorting and the crowding distance strategy [14]. The pseudocode of NSGA-II-ECFE is summarized in Algorithm 1.

2) MOEA/D-ECFE: In MOEA/D-ECFE, it decomposes the multiobjective clustering problem with those four objective functions into NP patient stratification clustering subproblems using the Tchebycheff method [15]. At each iteration, a population with NP subproblems is initialized and assigned a set of uniformly distributed weight vectors λ over the weight space. For each subproblem, an offspring solution is generated by employing the uniform crossover operator [29] on two random subproblems chosen from the current population. Then, the neighborhood-biased mutation operator [29] is applied to each basic clustering of the ensemble in the offspring solution to enhance the exploitation ability of MOEA/D-ECFE. Each subproblem is optimized using the information from its NS nearest subproblems. Finally, a new population with NP subproblems is produced. The pseudocode of MOEA/D-ECFE is presented in Algorithm 2.

Algorithm 1: Pseudocode of NSGA-II-ECFE

```
Input: (1) Ensemble size ($M$); (2) Population size ($NP$); (3) Number of
        fitness evaluations ($FES$);
Output: (1) Normalized Mutual Information (NMI); (2) Adjusted Rand
        Index (ARI);
Initialize the population $P = \{p_1, p_2, ..., p_i, ..., p_{NP}\}$;
$p_i = \left\{\Pi^{(i)}, \Psi^{(i)}\right\} =$
$\left\{\pi_1^{(i)}, \pi_2^{(i)}, ..., \pi_M^{(i)}, \alpha^{(i)}, \beta_1^{(i)}, \beta_2^{(i)}, \beta_3^{(i)}\right\}, i \in \{1, 2, ..., NP\}$;
$F(P) \leftarrow ECFE(P)$;
Perform fast non-dominated sorting to calculate the rank of individuals in $P$;
Calculate the crowding distance;
while the stopping criterion is not satisfied do
    Generate a mating pool;
    for $i = 1 \rightarrow NP$ do
        if $i < NP$ then
            $j \leftarrow i, k \leftarrow i + 1$;
        else
            $j \leftarrow 1, k \leftarrow NP$;
        Select two individuals $p_j$ and $p_k$ from the mating pool;
        Perform crossover and mutation operators on each basic
        clustering in $\Pi^{(j)} \in p_j$ and $\Pi^{(k)} \in p_k$ to obtain a new
        ensemble $\Pi^{(new)}$;
        $\Psi^{(new)} \leftarrow \Psi^{(i)}$;
        $p_{new} \leftarrow \left\{\Pi^{(new)}, \Psi^{(new)}\right\}$;
        $F(p_{new}) \leftarrow ECFE(p_{new})$;
        $P_{new} \leftarrow \{P, p_{new}\}$;
    Perform fast non-dominated sorting on the new population $P_{new}$ with
    $2NP$ individuals;
    Calculate the crowding distance;
    Choose the population with top $NP$ individuals for the next iteration;
Produce the Pareto set $\widehat{PS}$ with all non-dominated consensus clusterings
under four objectives;
Return the best NMI ($\max_{i\in\widehat{PS}} NMI_i$) and the corresponding ARI;
```

D. Time Complexity

This section focused on the time complexity of NSGA-II-ECFE and MOEA/D-ECFE. For each iteration of them, NSGA-II costs $O(I \times NP^2)$ and the time complexity of MOEA/D is $O(I \times NP \times NS)$ [30], where I is the number of objective functions, NP is the number of individuals in the population, and NS is the size of neighbors. Considering the computation of ECFE, it costs the worst time $O(n^2 \times D \times NP)$ [31], where n is the number of samples in the given dataset, D is the number of genes after the dimensionality reduction. Therefore, the overall worst time complexity of NSGA-II-ECFE and MOEA/D-ECFE per iteration is $O(I \times NP^2 + n^2 \times D \times NP)$ and $O(I \times NP \times NS + n^2 \times D \times NP)$ respectively.

IV. Experiments and Results

A. Data Collection

We collect fifty-five synthetic datasets [10] based on a real human transcriptional regulation network. They are generated as follows:

$$\begin{aligned} F_i^{mRNA}(\mathbf{x}, \mathbf{y}) &= \frac{dx_i}{dt} = m_i \cdot f_i(\mathbf{y}) - \lambda_i^{mRNA} \cdot x_i \\ F_i^{Prot}(\mathbf{x}, \mathbf{y}) &= \frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{Prot} \cdot y_i \end{aligned} \tag{13}$$

Algorithm 2: Pseudocode of MOEA/D-ECFE

```
Input: (1) Ensemble size ($M$); (2) Neighbor size ($NS$); (3) Number of
       fitness evaluations ($FES$);
Output: (1) Normalized Mutual Information (NMI); (2) Adjusted Rand
        Index (ARI);
Initialization:
Coefficient vectors $\lambda = \left\{\lambda^1, \lambda^2, ...., \lambda^{NP}\right\}$, where
$\lambda^i = \{\lambda_1^i, \lambda_2^i, \lambda_3^i, \lambda_4^i\}$, $i \in \{1, 2, ..., NP\}$;
The population $P = \{p_1, p_2, ..., p_i, ..., p_{NP}\}$ with $NP$ subproblems;
$p_i = \left\{\Pi^{(i)}, \Psi^{(i)}\right\} =$
$\left\{\pi_1^{(i)}, \pi_2^{(i)}, ..., \pi_M^{(i)}, \alpha^{(i)}, \beta_1^{(i)}, \beta_2^{(i)}, \beta_3^{(i)}\right\}, i \in \{1, 2, ..., NP\}$;
For each subproblem $p_i$, find $NS$ closest weight vectors
$B_i = \{i_1, i_2, ..., i_{NS}\}$;
$F(P) \leftarrow ECFE(P)$;
The reference point $z^* = \{z_1^*, z_2^*, z_3^*, z_4^*\}$ with the ideal objective value
found so far for each objective function;
Iteration:
while the stopping criterion is not satisfied do
    for $i$ = 1 $\rightarrow NP$ do
        Randomly select two subproblems $p_j$ and $p_k$ from $P$;
        Perform crossover and mutation operators on $\Pi^{(j)} \in p_j$ and
        $\Pi^{(k)} \in p_k$ to obtain a new ensemble $\Pi^{(new)}$;
        $\Psi^{(new)} \leftarrow \Psi^{(i)}$;
        $p_{new} \leftarrow \left\{\Pi^{(new)}, \Psi^{(new)}\right\}$;
        Update the reference point $z^*$;
        for $j \in B_i$ do
            if $g^{te}(p_{new}|\lambda^i) < g^{te}(p_j|\lambda^i)$ then
                $p_j \leftarrow p_{new}$;
                $F(p_j) \leftarrow F(p_{new})$;
            else
                Generate a new parameter vector $\Psi^{(j)}$ for $p_j$;
Produce the Pareto set $\widehat{PS}$ with all non-dominated consensus clusterings
under four objectives;
Return the best NMI ($\max_{i \in \widehat{PS}} NMI_i$) and the corresponding ARI;
```

where m_i represents the maximum transcription rate, r_i is the translation rate, λ_i^{mRNA} and λ_i^{Prot} are the mRNA and protein deterioration rates, **x** and **y** are the mRNA and protein concentration level vectors, and $f_i(\cdot)$ is the activation function of the i-th gene. The characteristics of those 55 synthetic datasets are summarized in Supplementary Table S1 [32]. It contains the noise level, the number of knock-out genes, samples, genes, and clusters. The number of knock-out genes is ranged from 100 to 500 and the noise level varies from 0 to 0.5.

In addition, all those thirty-five patient stratification datasets are obtained from [10]. Supplementary Table S2 provides their details [33] including the data source, the number of samples, clusters, and genes. The minimum number of samples is 22 and the maximum number of samples is 248; the number of clusters ranges from 2 to 14; the number of genes varies in [85, 4553].

Finally, we adopt five cancer-related single-cell RNA datasets [34] to demonstrate the performance of the proposed algorithm in real-world applications.

B. Evaluation Metrics

Two widely-used evaluation metrics, called Normalized Mutual Information (*NMI*) [35] and Adjusted Rand Index (*ARI*) [7], are used to evaluate the clustering performance. They can provide a sound indication of the similarities between the predicted and ground truth label. The clustering results with higher values indicate better clusterings.

NMI measures the shared information between two clustering results and varies from 0 to 1; *ARI* provides the agreement between two clustering results and ranges from -1 to 1. Given π_e the predicted label and π_t the ground truth label, they are defined as follows:

$$NMI(\pi_e, \pi_t) = \frac{\sum_{i=1}^{n_e} \sum_{j=1}^{n_t} n_{ij} \log \frac{n_{ij} n}{n_e^i n_t^j}}{\sqrt{(\sum_{i=1}^{n_e} n_e^i \log \frac{n_e^i}{n})(\sum_{j=1}^{n_t} n_t^j \log \frac{n_t^j}{n})}} \tag{14}$$

$$ARI(\pi_e, \pi_t) = \frac{\sum_{i=1}^{n_e} \sum_{j=1}^{n_t} \binom{n_{ij}}{2} - \sum_{i=1}^{n_e} \binom{n_e^i}{2} \cdot \sum_{j=1}^{n_t} \binom{n_t^j}{2} / \binom{n}{2}}{\sum_{i=1}^{n_e} \binom{n_e^i}{2} / 2 + \sum_{j=1}^{n_t} \binom{n_t^j}{2} / 2 - \sum_{i=1}^{n_e} \binom{n_e^i}{2} \cdot \sum_{j=1}^{n_t} \binom{n_t^j}{2} / \binom{n}{2}} \tag{15}$$

where n_e, n_t are the cluster numbers in π_e and π_t, respectively. n_e^i is the number of samples in cluster i of π_e, n_t^j is the number of samples in cluster j of π_t, and n_{ij} is the intersection sample size between clusters i and j.

To measure the overall performance of those baseline methods over those patient stratification datasets, we adopt an average performance score [10], which can be formulated as follows:

$$Avg(ALGO_i) = \frac{1}{d} \sum_{j=1}^{d} \frac{V(Dataset_j, ALGO_i)}{max_i V(Dataset_j, ALGO_i)} \tag{16}$$

where $V(Dataset_j, ALGO_i)$ denotes the evaluation metric (*NMI* or *ARI*) of the i-th algorithm $ALGO_i$ on the j-th dataset $Dataset_j$, d denotes the total number of benchmark datasets.

C. Parameter Settings

In the proposed algorithm, the ensemble size (M) is set to 20 and the number of genes after dimensionality reduction (D) is set to 300. In particular, for NSGA-II-ECFE, we set the population size $NP = 200$. For MOEA/D-ECFE, the population size NP is equal to the number of weight vectors C_{H+I-1}^{I-1}, in which I is the number of objectives 4 and H is 7. The size of neighbors NS for each weight vector is 2. The discussions about those parameters are presented in Supplementary Section I. The parameter α of the proposed algorithms NSGA-II-ECFE and MOEA/D-ECFE is the binary number, which is discussed in Section III (A). Meanwhile, in Tables S4-S15, Table S18, and Tables S20-S25, the parameter α of the proposed algorithms NSGA-II-ECFE and MOEA/D-ECFE is the binary number. To conduct a fair comparison, the number of fitness evaluations (FES) is taken as the stopping criterion. We set FES as 1000 for each dataset [36]. Meanwhile, the average *NMI* and *ARI* are provided over 30 independent runs on each patient stratification dataset to exclude the factor of *getting lucky occasionally*.

For the single clustering algorithms, the following experimental settings are adopted.

- For SC, the similarity graph with the size of $n \times n$ is constructed by the k-nearest neighbor graph method [17].
- For DBSCAN, *eps* is set to 0.5, and *minPts* is set to 5 [37].

- For CDPs, the cutoff distance threshold (d_c) is set to 2, and the density is computed using the Gaussian kernel [19].

To make a fair comparison, for those twelve ensemble clustering algorithms, the base clusterings are generated by k-means clustering method, the number of base clusterings (ensemble size) is set to 20, the number of clusters in each base clustering is selected randomly from [2,$\sqrt{n}$] [20], where n is the number of samples in the patient stratification dataset, and the low variance method [16] is adopted to reduce the dimension for all ensemble clustering baseline algorithms.

D. Baseline Methods

Several methods are adopted to demonstrate the performance of NSGA-II-ECFE and MOEA/D-ECFE. From the clustering perspective, we compared them against seven clustering methods, including agglomerative hierarchical clustering with average-linkage (AL) [38], single-linkage (SL) [38], complete-linkage (CL) [38], KM [18], SC [17], density-based spatial clustering with noise (DBSCAN) [37], and CDPs [19]. The reason of choosing those clustering approaches is that, AL, SL, CL, and KM are simple clustering methods usually applied to analyze data; SC is a graph theory-based clustering algorithm; DBSCAN is an effective clustering method based on density; and CDPs uses the density peaks to discover the cluster centers.

From the ensemble clustering perspective, we compared the proposed algorithms with twelve ensemble clustering methods, namely, linked-based cluster ensemble (LCE) [39], cluster-based similarity partitioning algorithm (CSPA) [40], hypergraph partitioning algorithm (HGPA) [40], meta-clustering algorithm (MCLA) [40], k-means-based consensus clustering (KCC) [41], spectral ensemble clustering (SEC) [42], entropy-based consensus clustering (ECC) [10], locally weighted ensemble clustering based on evidence accumulation (LWEA) [43], locally weighted ensemble clustering based on graph partitioning (LWGP) [43], ultra-scalable ensemble clustering (U-SENC) [31], ensemble clustering by propagating cluster-wise similarities based on hierarchical clustering (ECPCS-HC) [23], and ensemble clustering by propagating cluster-wise similarities based on meta-clustering (ECPCS-MC) [23]. LCE employs a linked-based algorithm to underly similarity assessment; CSPA, HGPA, and MCLA are ensemble clustering algorithms based on graph partitioning; KCC is a consensus clustering method based on k-means clustering; SEC is a spectral ensemble clustering algorithm using co-association matrix; ECC employs the entropy-based utility function to merge the basic clustering into a consensus clustering; LWEA and LWGP utilize local weighting strategy based on two different consensus functions; U-SENC is an ensemble clustering framework that integrates multiple clusters generated by the ultra-scalable spectral clustering; ECPCS-HC and ECPCS-MC are two ensemble clustering approaches based on fast propagation of cluster-wise similarities via random walks with two different consensus functions. Moreover, the time and space complexity of those different ensemble clustering methods are summarized in Supplementary Table S3.

From the multiobjective perspective, we compared NSGA-II-ECFE and MOEA/D-ECFE with two multiobjective clustering algorithms, including MOCDP [8] and MOSC [9]. They are multiobjective clustering algorithms based on the CDPs clustering and the spectral clustering, respectively.

E. Synthetic Datasets

In this section, we compare the performance of NSGA-II-ECFE and MOEA/D-ECFE with those seven clustering methods and twelve ensemble clustering methods on 55 synthetic datasets. To conduct a fair comparison, we run each algorithm 30 times on each dataset. Meanwhile, the performance is evaluated by the average *NMI* and *ARI* score. The experimental results of all those clustering algorithms are summarized in Fig. 3. From Fig. 3, it can be found that NSGA-II-ECFE performs better than other clustering algorithms for all the datasets while MOEA/D-ECFE is superior to other competitive methods for most datasets. For 100 knock-out genes, at the noise level of 0.45, NSGA-II-ECFE can provide *NMI* and *ARI* improvement over MOEA/D-ECFE at about 4.7% and 3.9%. In particular, compared with AL and SL, the proposed algorithms can provide high-quality results at most of the noise levels by a large margin.

For ensemble clustering methods, observing the results summarized in Fig. 4, we can find that NSGA-II-ECFE and MOEA/D-ECFE are better than or equal to those methods on most datasets. For 200 knock-out genes, at the noise level 0.35, CSPA is superior to NSGA-II-ECFE; while for 400 knock-out genes, at the noise level 0.45, NSGA-II-ECFE is slightly inferior to other ensemble clustering algorithms except LCE, SEC, U-SENC, and ECPCS-HC. For 500 knock-out genes, at the noise level 0.45, CSPA surpasses NSGA-II-ECFE. For the rest datasets, NSGA-II-ECFE achieves promising solutions. It is worth noting that all those methods enable to group the datasets at very small noise levels for the low perturbation with the human transcriptional regulation network. Besides, Supplementary Fig. S1 shows the average scores of those methods to measure their overall performance concerning *NMI* and *ARI* respectively. It can be observed that the proposed algorithms outperform all those comparative algorithms and NSGA-II-ECFE is slightly competitive to MOEA/D-ECFE. Therefore, it is empirically validated that the proposed algorithms have significant advantages on those fifty-five synthetic datasets in a robust manner.

F. Patient Stratification Datasets

To further demonstrate the advantages of NSGA-II-ECFE and MOEA/D-ECFE, we compare them with seven clustering methods and twelve ensemble clustering methods on 35 patient stratification datasets. Each method runs 30 times on each dataset for a fair comparison. The average performance of each algorithm is measured by the average *NMI* and *ARI* score. For statistically rigorous comparisons, Friedman test [44] is used to show the average ranking of all algorithms. In addition, the paired Wilcoxon test is calculated to show the statistical difference between the lowest-ranked algorithm and other algorithms with a significant level 0.05. H_1 denotes there

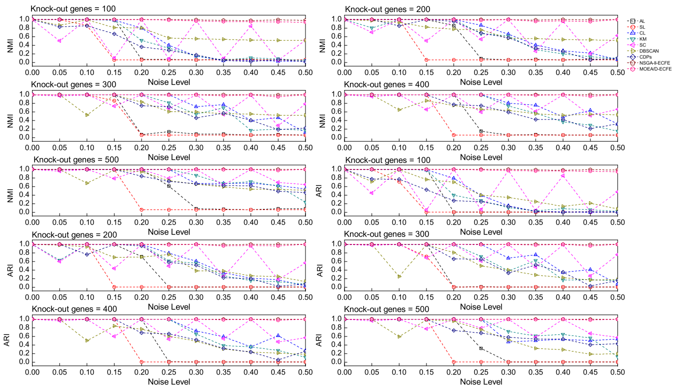


Fig. 3: Comparison performance of different clustering algorithms on those fifty-five synthetic datasets

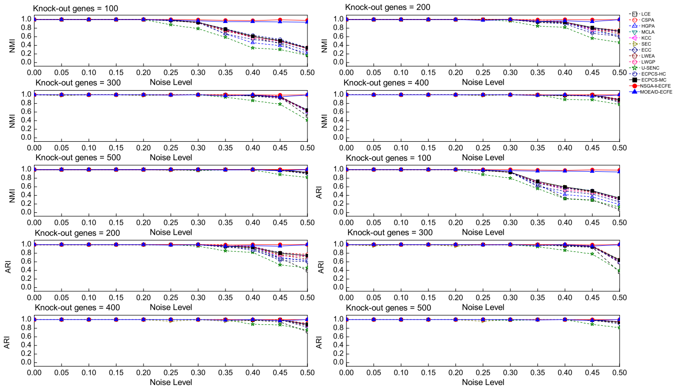


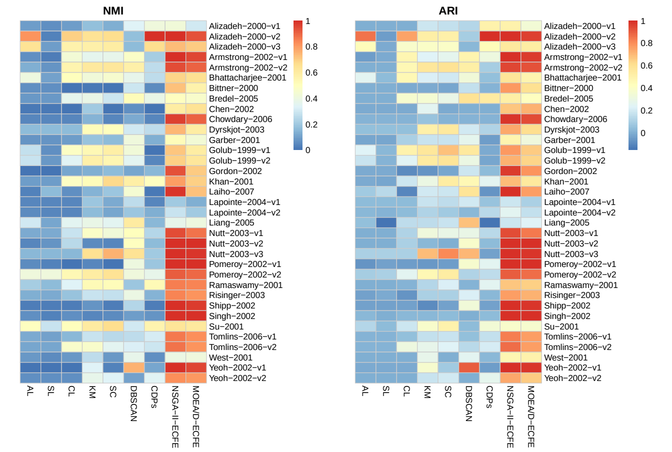
Fig. 4: Comparison performance of different ensemble clustering algorithms on those fifty-five synthetic datasets

is significant difference between them while H_0 represents the algorithms are statistically equivalent to each other.

1) Comparison with Clustering Methods: Considering those seven clustering methods, the comparative experimental results measured by *NMI* and *ARI* are summarized in Supplementary Table S4 and Table S5. The last three rows list the mean ranks and the Wilcoxon test results of those clustering methods. Meanwhile, Fig. 5 illustrates *NMI* and *ARI* results of those clustering methods on 35 patient stratification datasets respectively.

In terms of *NMI*, from Supplementary Table S4, it can be observed that NSGA-II-ECFE and MOEA/D-ECFE obtain the best *NMI* for 28 and 5 datasets out of 35 datasets respectively. For Alizadeh-2000-v2, CDPs can provide the best *NMI*; for Su-2001, SC obtains a slightly better *NMI* result than NSGA-II-ECFE. Meanwhile, for Bredel-2005, Lapointe-2004-v1, and Liang-2005, DBSCAN can achieve the best *NMI* results. NSGA-II-ECFE and MOEA/D-ECFE generally reach the best average *NMI* result across 35 datasets with the lowest ranks. In addition, NSGA-II-ECFE and MOEA/D-ECFE are significantly different from other clustering algorithms. It can be validated that the proposed algorithms, particularly NSGA-II-ECFE, have great efficacy in clustering patient stratification data.

In terms of *ARI*, from Supplementary Table S5, NSGA-II-ECFE and MOEA/D-ECFE have a relatively low rank among all the methods with the performance that is significantly different from other algorithms. For AL, SL, CL, and KM, they cannot obtain the best *ARI* results for any of those patient stratification datasets; for SC, it can obtain the best *ARI* results on Su-2001; for DBSCAN, it achieves the best *ARI* on Bredel-2005, Lapointe-2004-v1, and Liang-2005; and for CDPs, it can provide the best *ARI* on Alizadeh-2000-v1 and Alizadeh-2000-v2.



(a)

(b)

Fig. 5: Comparison performance of different clustering algorithms on thirty-five patient stratification datasets. The performance is measured by *NMI* in (a) and *ARI* in (b) in heatmap colors.

2) Comparison with Other Ensemble Clustering Algorithms: In this section, we will compare our proposed algorithm with other ensemble clustering algorithms on 35 patient stratification datasets. The experimental results are summarized in Supplementary Table S6 and Table S9 respectively.

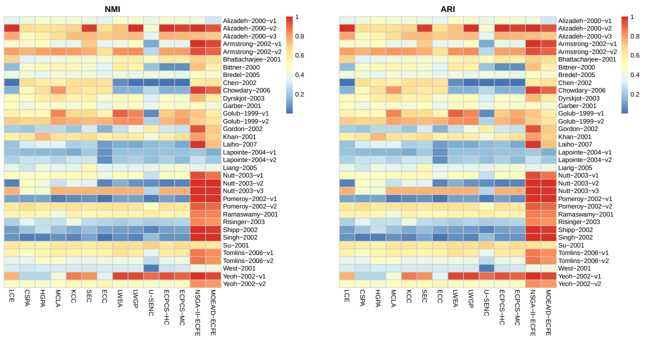
Regarding *NMI*, from Supplementary Table S6, NSGA-II-ECFE can provide the best *NMI* results on most datasets except eleven datasets. In particular, NSGA-II-ECFE can achieve perfect clustering on Armstrong-2002-v1, Nutt-2003-v2, Nutt-2003-v3, Pomeroy-2002-v1, Shipp-2002, and Singh-2002. While MOEA/D-ECFE can perform the same on three patient stratification datasets including Nutt-2003-v2, Pomeroy-2002-v1, and Singh-2002. The reason may be that each of those datasets has only two clusters with a small number of samples with a clear structure. Moreover, since our proposed algorithm is a multiobjective evolutionary algorithm, it applied multiple objective functions to capture diverse characteristics of those datasets, resulting in good performance. To further analyze such phenomenon, we use two other external evaluation metrics including clustering accuracy and purity, and two other internal evaluation metrics including DB and Dunn, to measure the clustering quality of different clustering algorithms on those six patient stratification datasets. Purity [45] is a point-level index that can measure the quality of the predicted clustering result, which can be defined as follows:

$$purity(\pi_e, \pi_t) = \frac{1}{n} \sum_{j=1}^{J} \underset{k}{max} \left| C_e^j \cap C_t^k \right| \qquad (17)$$

where $\pi_e = \left\{ C_e^1, C_e^2, ..., C_e^j, ..., C_e^J \right\}$ is the predicted clustering result with J clusters, $\pi_t = \left\{ C_t^1, C_t^2, ..., C_t^k, ..., C_t^K \right\}$ is the set of K classes in the ground truth label, and n is the number of samples in the dataset. Each cluster is assigned to the class that is most frequent in the cluster. Normally, larger purity values indicate that the predicted clustering result has better quality. The clustering accuracy is equivalent to the purity [46]. DB and Dunn are defined in Eq. (9) and Eq. (10).

The results are summarized in Supplementary Table S7 and Table S8 on those six patient stratification datasets respectively. From Supplementary Table S7 and Table S8, in terms of *NMI*, *ARI*, clustering accuracy, and purity, it can be found that NSGA-II-ECFE can achieve the best results among those ensemble clustering algorithms and multiobjective clustering algorithms, which indicates that the proposed algorithm reveals significant advantages over all other methods. Besides, we have added two internal evaluation metrics including DB and Dunn to analyse this phenomenon. Since the truth labels are available for all datasets, we calculate the absolute values of DB and Dunn based on the resulted cluster labels and ground truth labels. DB_{algo} and Dunn_{algo} denote the DB and Dunn which are computed on the resulted cluster labels obtained by a given algorithm while DB_{truth} and Dunn_{truth} denote the DB and Dunn which are computed on the ground truth labels. Therefore, a small value indicates a good clustering result close to the ground truth result. We can observe that the performance of NSGA-II-ECFE is superior to other clustering algorithms. In particular, the DB_{algo} and Dunn_{algo} of NSGA-II-ECFE are equal to DB_{truth} and Dunn_{truth} on those six datasets while those of MOEA/D-ECFE are equal to DB_{truth} and Dunn_{truth} on three patient stratification datasets including Nutt-2003-v2, Pomeroy-2002-v1, and Singh-2002. Therefore, we can conclude that the proposed NSGA-II-ECFE, can achieve promising results on those datasets. NSGA-II-ECFE has the lowest mean rank across those 35 datasets. Meanwhile, from the statistical results by the paired Wilcoxon test, we can find that there is significant difference between the proposed algorithm and other ensemble clustering algorithms. In addition, Fig. 6 (a) shows the clustering performance of those algorithms evaluated by *NMI*.

Regarding *ARI*, from Supplementary Table S9, MOEA/D-ECFE and NSGA-II-ECFE outperform other ensemble clustering algorithms. Meanwhile, the average *ARI* score increasing of NSGA-II-ECFE over LCE, CSPA, HGPA, MCLA, KC-C, SEC, ECC, LWEA, LWGP, U-SENC, ECPCS-HC, and ECPCS-MC, is 46.7%, 45.2%, 42.1%, 38.9%, 38.2%, 34.4%, 49.9%, 43.2%, 41.4%, 63%, 46.5% and 46.7% respectively. For LCE, it obtains the best *ARI* on 3 datasets including Alizadeh-2000-v2, Bhattacharjee-2001, and Garber-2001. For SEC and LWEA, they can achieve the best *ARI* on 2 out of 35 patient stratification datasets. For HGPA, KCC, LWGP, U-SENC, ECPCS-HC, and ECPCS-MC, they provide the best *ARI* for one dataset. For the rest ensemble clustering algorithms, they cannot achieve the best *ARI* on any datasets. Furthermore, NSGA-II-ECFE provides the lowest rank among all algorithms. By observing the statistical results of the paired Wilcoxon test, NSGA-II-ECFE and MOEA/D-ECFE can accurately group the patient stratification data in a statistically significant manner. Besides, Fig. 6 (b) also depicts their performance measured by *ARI* clearly and Supplementary Fig. S2 shows the ANOVA analysis to demonstrate the robustness of NSGA-II-ECFE. Based on those observations, we can conclude that the proposed algorithms can provide better performance than other methods, which can be served as an effective technique for clustering patient stratification data.



(a)

(b)

Fig. 6: Comparison performance of different ensemble clustering algorithms on thirty-five patient stratification datasets. The performance is measured by *NMI* in (a) and *ARI* in (b) in heatmap colors.

G. Multiobjective Optimization Methodology Comparisons

To investage the effectiveness of the proposed algorithms from the multiobjective perspective, we compare the proposed NSGA-II-ECFE and MOEA/D-ECFE against two multiobjective clustering algorithms, including MOCDP and MOSC on 35 patient stratification datasets. Each algorithm has been independently run for 30 times on each dataset. The experimental results are measured by *NMI* and *ARI* and summarized in Supplementary Table S10 and Table S11 respectively.

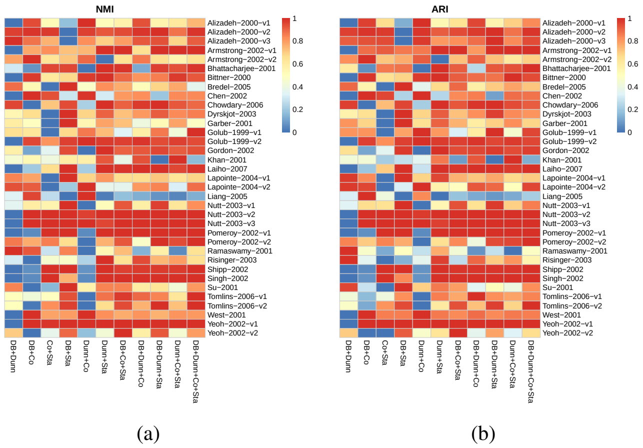
From Supplementary Table S10, we can observe that NSGA-II-ECFE is the best-performing algorithm with the lowest rank among all algorithms. Compared with MOCDP, MOSC, and MOEA/D-ECFE, it can be found that NSGA-II-ECFE can provide better *NMI* results on 25, 19, and 29 patient stratification datasets respectively. In addition, from Supplementary Table S11, our proposed NSGA-II-ECFE can obtain a better average *ARI* score than other three algorithms. For MOCDP, it achieves the best *ARI* on two datasets including Alizadeh-2000-v2 and Bredel-2005 while it is inferior to the proposed algorithms on most datasets. MOSC achieves the best *ARI* on 15 out of 35 datasets while NSGA-II-ECFE increases the average *ARI* score across those datasets by 5.7% over it. As evidenced by those experimental results, we claim that our proposed algorithms, in particular NSGA-II-ECFE, is superior to other multiobjective clustering algorithms in stratifying those patient stratification datasets into subtypes.

H. Extended Analysis and Comparisons

1) Different Objective Function Subsets: In this section, to demonstrate the effectiveness of different objective function combinations for the proposed algorithm NSGA-II-ECFE, we compare NSGA-II-ECFE under 11 different combinations of objective functions on 35 patient stratification datasets. Each objective function combination is chosen from those four objective functions including DB, Dunn, cohesion, and stability. The experimental results are tabulated in Supplementary Tables S12-S15. The last row of each table summarizes the average *NMI* (*ARI*) score to evaluate the overall performance of each algorithm over those 35 datasets. Moreover, Fig. 7 visualizes those *NMI* and *ARI* results of NSGA-II-ECFE under different objective functions subsets.

In terms of *NMI*, from Supplementary Table S12, Table S14 and Fig. 7 (a), it is pointed out that NSGA-II-ECFE outperforms other algorithms under different two objective functions subsets and three objective functions subsets on most datasets. Besides, NSGA-II-ECFE can provide the best average *NMI* score in all the compared algorithms, which indicates the effectiveness of NSGA-II-ECFE under four objective functions. In terms of *ARI*, from Supplementary Table S13, Table S15, and Fig. 7 (b), it can be observed that NSGA-II-ECFE can yield better performance than other algorithms under different objective function combinations. The largest and least average *ARI* score improvements of NSGA-II-ECFE over the other algorithms are 31.75% and 1.92% respectively. Based on the analysis, it can be demonstrated that NSGA-II-ECFE under those four objective functions exhibits competitive edges over others on most patient stratification datasets.

Besides, the DB and Dunn tendency under the number of fitness evaluations (FES) of those thirty-five patient stratification datasets provided by NSGA-II-ECFE are summarized in Supplementary Fig. S7. As shown in Supplementary Fig. S7, it can be found that for most patient stratification datasets, the curve of DB provides a downward trend and the curve of Dunn has an upward trend. It conforms to the minimization and maximization optimization of DB and Dunn respectively, demonstrating the effectiveness of the proposed algorithm.



(a) (b)

Fig. 7: Comparison performance of NSGA-II-ECFE under different objective functions subsets on thirty-five patient stratification datasets. The performance is measured by *NMI* in (a) and *ARI* in (b) in heatmap colors.

2) Dimensionality Reduction: To demonstrate the effect of the dimensionality reduction, we compare NSGA-II-ECFE with NSGA-II-ECFE without dimensionality reduction (NSGA-II-ECFE$_{noDR}$) in this section. The performance is measured by the average *NMI* over 30 runs on 35 patient stratification datasets. The experimental results of each algorithm on each dataset and the average *NMI* score through those datasets are summarized in Fig. 8. As shown in Fig. 8, the dimensionality reduction contributes to enhancing the overall performance of the proposed algorithms.

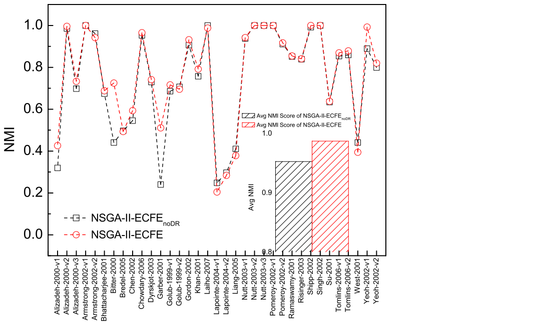


Fig. 8: Comparison performance of NSGA-II-ECFE with dimensionality reduction and NSGA-II-ECFE without dimensionality reduction by average *NMI* across thirty-five patient stratification datasets.

3) Effect of Ensemble Construction Method: In this section, to demonstrate the effectiveness of the proposed ensemble construction method, we compare NSGA-II-ECFE based on the proposed ensemble construction method with NSGA-II-ECFE based on other two ensemble construction methods. The first comparative method is to construct all clusterings by k-means clustering method; the other method is only by the locus-based adjacency method. They are named NSGA-II-ECFE$_1$ and NSGA-II-ECFE$_2$ respectively. The comparative experiment is performed on 35 patient stratification datasets over 30 runs. The results measured by the average *NMI* are summarized in Fig. 9. As depicted in Fig. 9, NSGA-II-ECFE with the proposed ensemble construction method provides 4% and 12.5% *NMI* improvement over NSGA-II-ECFE$_1$ and NSGA-II-ECFE$_2$ respectively. Therefore, we can conclude that our proposed ensemble construction method are synergistic with the proposed algorithm.

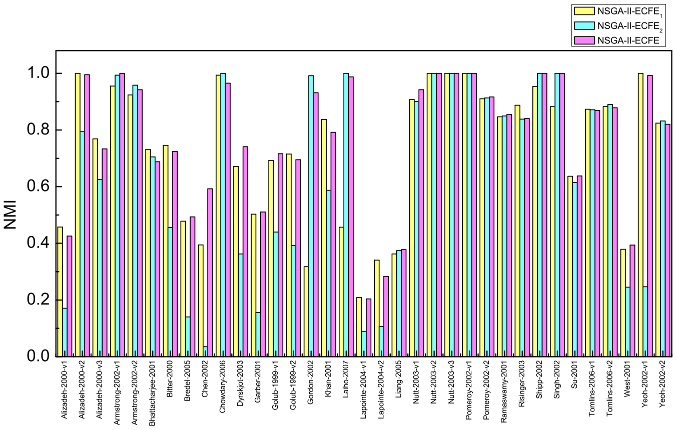


Fig. 9: Performance comparison of NSGA-II-ECFE with three ensemble construction methods by average *NMI* across thirty-five patient stratification datasets.

4) Comparison with Different Types of Co-association Matrix: This section is dedicated to demonstrate the performance of the hybrid co-association matrix (HCM) in ECFE by comparing it with NSGA-II-ECFE using the evidence accumulation co-association matrix (NSGA-II-ECFE$_{EACM}$) and NSGA-II-ECFE using the enhanced co-association matrix (NSGA-II-ECFE$_{ECM}$) on those 35 patient stratification datasets. Fig. 10 illustrates the average *NMI* results over 30 runs for each dataset. As shown in Fig. 10, it can conclude that NSGA-II-ECFE with HCM can yield better performance than other co-association matrices on those datasets. The proposed hybrid co-association matrix is more appropriate for ECFE to stratify patients into subtypes than other two types of co-association matrix. In particular, for Pomeroy-2002-v1, Shipp-2002, and Singh-2002, NSGA-II-ECFE with the hybrid co-association matrix shows its great superiority to other different types of co-association matrix by achieving over 70% *NMI* improvement.

Furthermore, for Shipp-2002, t-distributed stochastic neighbor embedding (t-SNE) [47] is implemented to project the similarity matrix into two dimensions to visualize different types of co-association matrix. The 2-D visualization of Shipp-2002 is depicted in Fig. 11. Notably, t-SNE visualizes the similarity matrix without the ground truth labels; the label is produced by NSGA-II-ECFE with the corresponding similarity matrix and formed in different colors to denote the results. From Fig. 11, it is shown that the proposed hybrid co-association matrix can represent the clustering structure of the patient stratification data more accurately compared with other two types of co-association matrix. Moreover, to demonstrate the effectiveness of the proposed HCM in ECFE further, we set α in HCM to a continuous value chosen from $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. The performance of those algorithms are measured by averaging *NMI* and *ARI* over 30 runs on 35 patient stratification datasets. The experimental results are summarized in Supplementary Table S16 and Table S17. As observed from those tables, we can observe that NSGA-II-ECFE with α setting to the binary number can provide better performance than the other algorithms on those patient stratification datasets in terms of *NMI* and *ARI*.

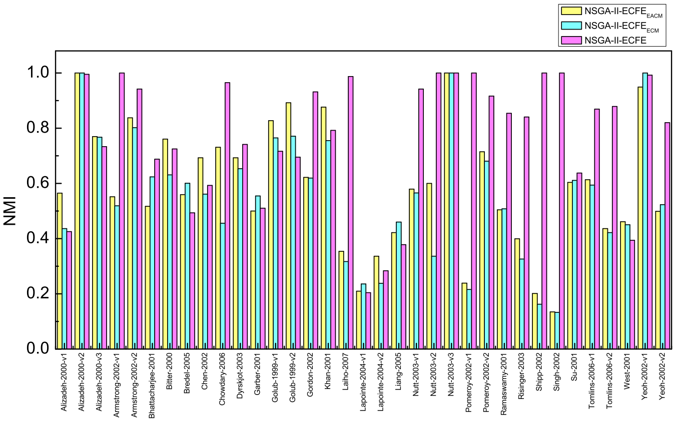
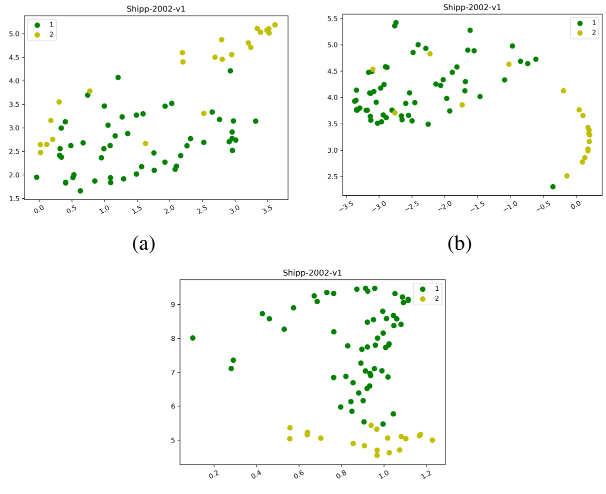


Fig. 10: Comparison performance of NSGA-II-ECFE with three different types of co-association matrix by average *NMI* across thirty-five patient stratification datasets.

5) Performance Comparison of Different Ensemble Clustering Methods on Runtimes: In this section, the runtime comparison experiment is conducted between the proposed algorithm and other ensemble clustering methods. We executed it on a PC with an i7-7500U CPU and 8GB of RAM in MATLAB. Supplementary Table S18 summarizes the runtime comparison results on thirty-five patient stratification datasets. From Supplementary Table S18, NSGA-II-ECFE and MOEA/D-ECFE cost long runtime on multiple datasets. That is because they are multiobjective iterative algorithms with hundreds of individuals. As mentioned in the time complexity analysis in Section III (D), the overall time complexity depends on the number of iterations, individuals, and genes as tabulated in Supplementary Table S2.



(a)

(b)

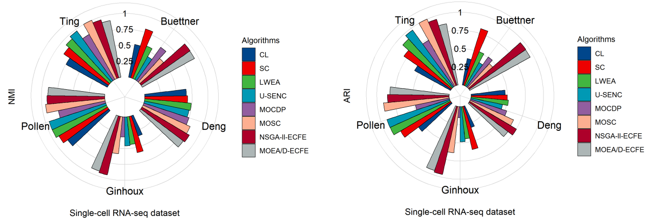
(c)

Fig. 11: The comparison of 2D visualization for Shipp-2002: (a) NSGA-II-ECFE$_{EACM}$; (b) NSGA-II-ECFE$_{ECM}$; (c) NSGA-II-ECFE. The axes are in arbitrary units in the projected space. Each point denotes a sample in Shipp-2002.

V. Application

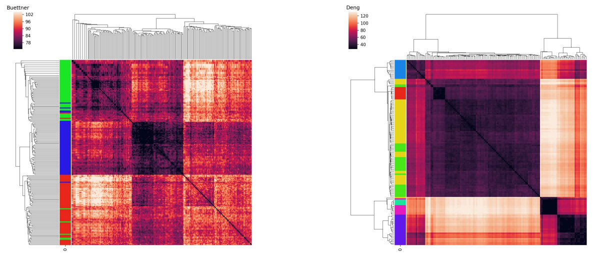
In this section, the proposed algorithms and other methods, including seven clustering algorithms, twelve ensemble clustering algorithms, and two multiobjective clustering algorithms, are applied to five cancer-related single-cell RNA-seq datasets to reveal the biological insights for NSGA-II-ECFE and MOEA/D-ECFE. Supplementary Table S19 summarizes the detailed dataset description, including the number of genes, single cells, and cancer subtypes. We measure the performance of each method on each dataset by *NMI* and *ARI*. The experimental results are tabulated from Supplementary Table S20 to Table S25. Besides, the comparison performances of different algorithms are shown in Fig. 12, in which we choose each top two algorithms from those three perspectives including the clustering, the ensemble clustering, and the multiobjective to compare with the proposed algorithms. As shown in Fig. 12 (a), NSGA-II-ECFE and MOEA/D-ECFE is superior to those algorithms on three single-cell RNA-seq datasets. For Pollen, NSGA-II-ECFE and MOEA/D-ECFE is slightly worse than LWEA, U-SENC, and MOSC; for Ting, MOSC performs better than the proposed algorithms; for the rest datasets, the proposed algorithms enhance the clustering performance by a certain margin, especially Ginhoux with 63% improvement mostly. Meanwhile, observing from Fig. 12 (b), it can be found that the proposed algorithms perform better than other methods on most datasets, except for Pollen and Ting.

To assess the clustering performance of the proposed algorithms in a visualization manner, we display the heatmap of Buettner and Deng with the estimated clustering from



(a) (b)

Fig. 12: Comparison performance of different algorithms measured by *NMI* (a) and *ARI* (b) on five cancer-related single-cell RNA-seq datasets.



(a) (b)

Fig. 13: Heatmap showing the expression of Buettner (a) and Deng (b) with the consensus clustering obtained by the NSGA-II-ECFE.

NSGA-II-ECFE in Fig. 13. As can be seen from Fig. 13, we find that the proposed algorithms can separate those single-cell RNA-seq data into several cell types by their gene expression profile, identifying the true data structure. In conclusion, the proposed algorithms are capable of learning cell-to-cell similarities from the gene expression data and capturing different representations of various single-cell datasets.

VI. Conclusion

In this study, we present a novel evolutionary multiobjective clustering with ensemble to cluster the patient stratification data in an effective and robust manner. It integrates an ensemble fitness evaluation (ECFE) method with an optimization framework (NSGA-II or MOEA/D) to generate the final consensus clustering. Four cluster validity indices, including DB, Dunn, cohesion, and stability, are employed to guide the evolution. In ECFE, to calculate the objective function fitness of the ensemble, a consensus clustering is generated from the ensemble. In order to produce the consensus clustering, a hybrid co-association matrix is proposed to represent the hierarchical structures of the patient stratification data, then, a suitable basic clustering algorithm is selected dynamically and employed on that similarity matrix. Several experiments are conducted to verify the performance of the proposed algorithm. The proposed algorithm provides significate advantages over other methods in terms of *NMI* and *ARI*, including seven clustering methods, twelve ensemble clustering methods, and two multiobjective clustering algorithms on fifty-five synthetic datasets and thirty-five patient stratification datasets. In addition, the time complexity and sensitivity analysis are analyzed to validate the performance of NSGA-II-ECFE and MOEA/D-ECFE from various perspectives. Moreover, we also applied them to analyze five cancer-related single-cell RNA-seq datasets. The results demonstrate that the proposed algorithms can identify cancer subtypes clearly. The source code of the proposed algorithm is available at https://github.com/wangyh082/ECFE.

In our future work, we plan to investigate the selection strategy in multiobjective optimization to choose the suitable solutions from those non-dominated consensus clusterings. Meanwhile, we would like to apply the proposed algorithms to other biological problems in the real world.

Acknowledgment

This research is supported by the National Natural Science Foundation of China under Grant No. 61603087, 61673331, and 62076109, and also funded by the Natural Science Foundation of Jilin Province under Grant No. 20190103006JH. The work described in this paper was substantially supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 11203217] and [CityU 11200218], onegrant from the Health and Medical Research Fund, the Food and Health Bureau, The Government of the Hong Kong Special Administrative Region [07181426], and the funding from Hong Kong Institute for Data Science (HKIDS) at City University of Hong Kong. The work described in this paper was partially supported by a grant from City University of Hong Kong (CityU 11202219).

References

[1] M. Steinbach, L. Ertöz, and V. Kumar, "The challenges of clustering high dimensional data," in *New directions in statistical physics*. Springer, 2004, pp. 273–309.

[2] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC bioinformatics*, vol. 15, no. 1, p. 37, 2014.

[3] C. Wang, R. Machiraju, and K. Huang, "Breast cancer patient stratification using a molecular regularized consensus clustering method," *Methods*, vol. 67, no. 3, pp. 304–312, 2014.

[4] S. Khakabimamaghani and M. Ester, "Bayesian biclustering for patient stratification," in *Biocomputing 2016: Proceedings of the Pacific Symposium*. World Scientific, 2016, pp. 345–356.

[5] K. Graim, T. T. Liu, A. S. Achrol, E. O. Paull, Y. Newton, S. D. Chang, G. R. Harsh, S. P. Cordero, D. L. Rubin, and J. M. Stuart, "Revealing cancer subtypes with higher-order correlations applied to imaging and omics data," *BMC medical genomics*, vol. 10, no. 1, p. 20, 2017.

[6] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "An interactive approach to multiobjective clustering of gene expression patterns," *IEEE Transactions on Biomedical Engineering*, vol. 60, no. 1, pp. 35–41, 2012.

[7] Y. Shi, Z. Yu, C. P. Chen, J. You, H.-S. Wong, Y. Wang, and J. Zhang, "Transfer clustering ensemble selection," *IEEE transactions on cybernetics*, 2018.

[8] X. Li and K.-C. Wong, "Evolutionary multiobjective clustering and its applications to patient stratification," *IEEE Transactions on Cybernetics*, no. 99, pp. 1–14, 2018.

[9] Y. Wang, Z. Ma, K.-C. Wong, and X. Li, "Nature-inspired multiobjective patient stratification from cancer gene expression data," *Information Sciences*, 2020.

[10] H. Liu, R. Zhao, H. Fang, F. Cheng, Y. Fu, and Y.-Y. Liu, "Entropy-based consensus clustering for patient stratification," *Bioinformatics*, vol. 33, no. 17, pp. 2691–2698, 2017.

[11] X. Yu, G. Yu, and J. Wang, "Clustering cancer gene expression data by projective clustering ensemble," *PloS one*, vol. 12, no. 2, 2017.

[12] L. I. Kuncheva and D. P. Vetrov, "Evaluation of stability of k-means cluster ensembles with respect to random initialization," *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, no. 11, pp. 1798–1808, 2006.

[13] S. J. Fodeh, C. Brandt, T. B. Luong, A. Haddad, M. Schultz, T. Murphy, and M. Krauthammer, "Complementary ensemble clustering of biomedical data," *Journal of biomedical informatics*, vol. 46, no. 3, pp. 436–443, 2013.

[14] K. Deb, A. Pratap, S. Agarwal, and T. Meyarivan, "A fast and elitist multiobjective genetic algorithm: Nsga-ii," *IEEE transactions on evolutionary computation*, vol. 6, no. 2, pp. 182–197, 2002.

[15] Q. Zhang and H. Li, "Moea/d: A multiobjective evolutionary algorithm based on decomposition," *IEEE Transactions on evolutionary computation*, vol. 11, no. 6, pp. 712–731, 2007.

[16] R. Silipo, I. Adae, A. Hart, and M. Berthold, "Seven techniques for dimensionality reduction," *White Paper by KNIME. com AG*, pp. 1–21, 2014.

[17] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[18] D. Steinley, "K-means clustering: a half-century synthesis," *British Journal of Mathematical and Statistical Psychology*, vol. 59, no. 1, pp. 1–34, 2006.

[19] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[20] X. Z. Fern and W. Lin, "Cluster ensemble selection," *Statistical Analysis and Data Mining: The ASA Data Science Journal*, vol. 1, no. 3, pp. 128–141, 2008.

[21] Y. Park and M. Song, "A genetic algorithm for clustering problems," in *Proceedings of the third annual conference on genetic programming*, vol. 1998, 1998, pp. 568–575.

[22] A. L. Fred and A. K. Jain, "Combining multiple clusterings using evidence accumulation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 27, no. 6, pp. 835–850, 2005.

[23] D. Huang, C.-D. Wang, H. Peng, J. Lai, and C.-K. Kwoh, "Enhanced ensemble clustering via fast propagation of cluster-wise similarities," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2018.

[24] U. Maulik and S. Bandyopadhyay, "Performance evaluation of some clustering algorithms and validity indices," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 24, no. 12, pp. 1650–1654, 2002.

[25] M. K. Pakhira, S. Bandyopadhyay, and U. Maulik, "Validity index for crisp and fuzzy clusters," *Pattern recognition*, vol. 37, no. 3, pp. 487–501, 2004.

[26] A. Mukhopadhyay, U. Maulik, and S. Bandyopadhyay, "A survey of multiobjective evolutionary clustering," *ACM Computing Surveys (CSUR)*, vol. 47, no. 4, pp. 1–46, 2015.

[27] C. Zhong, L. Hu, X. Yue, T. Luo, Q. Fu, and H. Xu, "Ensemble clustering based on evidence extracted from the co-association matrix," *Pattern Recognition*, vol. 92, pp. 93–106, 2019.

[28] H. Chang and D.-Y. Yeung, "Robust path-based spectral clustering," *Pattern Recognition*, vol. 41, no. 1, pp. 191–203, 2008.

[29] J. Handl and J. Knowles, "An evolutionary approach to multiobjective clustering," *IEEE transactions on Evolutionary Computation*, vol. 11, no. 1, pp. 56–76, 2007.

[30] A. Gupta, S. Datta, and S. Das, "Fuzzy clustering to identify clusters at different levels of fuzziness: An evolutionary multiobjective optimization approach," *IEEE transactions on cybernetics*, 2019.

[31] D. Huang, C.-D. Wang, J. Wu, J.-H. Lai, and C. K. Kwoh, "Ultra-scalable spectral clustering and ensemble clustering," *IEEE Transactions on Knowledge and Data Engineering*, 2019.

[32] H. Y. Chang, D. S. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie, H. Dai, Y. D. He, L. J. van't Veer, H. Bartelink *et al.*, "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival," *Proceedings of the national academy of sciences*, vol. 102, no. 10, pp. 3738–3743, 2005.

[33] M. C. de Souto, I. G. Costa, D. S. de Araujo, T. B. Ludermir, and A. Schliep, "Clustering cancer gene expression data: a comparative study," *BMC bioinformatics*, vol. 9, no. 1, p. 497, 2008.

[34] S. Park and H. Zhao, "Spectral clustering based on learning similarity matrix," *Bioinformatics*, vol. 34, no. 12, pp. 2069–2076, 2018.

[35] X. Li and K.-C. Wong, "Multiobjective patient stratification using evolutionary multiobjective optimization," *IEEE journal of biomedical and health informatics*, vol. 22, no. 5, pp. 1619–1629, 2017.

[36] X. Li, S. Ma, and K.-C. Wong, "Evolving spatial clusters of genomic regions from high-throughput chromatin conformation capture data," *IEEE transactions on nanobioscience*, vol. 16, no. 6, pp. 400–407, 2017.

[37] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise." in *Kdd*, vol. 96, no. 34, 1996, pp. 226–231.

[38] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, no. 11, pp. 1370–

1386, 2004.

[39] N. Iam-On, T. Boongoen, and S. Garrett, "Lce: a link-based cluster ensemble method for improved gene expression data analysis," *Bioinformatics*, vol. 26, no. 12, pp. 1513–1519, 2010.

[40] A. Strehl and J. Ghosh, "Cluster ensembles—a knowledge reuse framework for combining multiple partitions," *Journal of machine learning research*, vol. 3, no. Dec, pp. 583–617, 2002.

[41] J. Wu, H. Liu, H. Xiong, J. Cao, and J. Chen, "K-means-based consensus clustering: A unified view," *IEEE transactions on knowledge and data engineering*, vol. 27, no. 1, pp. 155–169, 2014.

[42] H. Liu, T. Liu, J. Wu, D. Tao, and Y. Fu, "Spectral ensemble clustering," in *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, 2015, pp. 715–724.

[43] D. Huang, C.-D. Wang, and J.-H. Lai, "Locally weighted ensemble clustering," *IEEE transactions on cybernetics*, vol. 48, no. 5, pp. 1460–1473, 2017.

[44] M. Friedman, "A comparison of alternative tests of significance for the problem of m rankings," *The Annals of Mathematical Statistics*, vol. 11, no. 1, pp. 86–92, 1940.

[45] E. Rendón, I. Abundez, A. Arizmendi, and E. M. Quiroz, "Internal versus external cluster validation indexes," *International Journal of computers and communications*, vol. 5, no. 1, pp. 27–34, 2011.

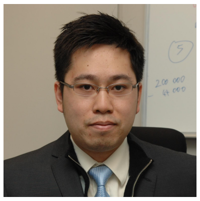
[46] M. Rezaei and P. Fränti, "Set matching measures for external cluster validity," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, no. 8, pp. 2173–2186, 2016.

[47] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of machine learning research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

Yunhe Wang is a PhD student with School of Information Science and Technology, Northeast Normal University, Changchun, China. She is now a visiting PhD student with School of Computer Science and Informatics, De Montfort University, Leicester, UK. Her current research interests include intelligent computation, machine learning.

Xiangtao Li (M'15) received the B.Eng. Degree, the M.Eng. and Ph.D. degrees in computer science from Northeast Normal University, Changchun, China in 2009, 2012, 2015, respectively.

He is currently a Professor with the School of Artificial Intelligence, Jilin University. He has published more than 50 research papers. His research interests include intelligent computation, evolutionary data mining, constrained optimization, bioinformatics, computational biology and interdisciplinary research.



Ka-Chun Wong received the BEng degree in computer engineering from United College, Chinese University of Hong Kong, in 2008. He received the MPhil degree from the same university in 2010 and the PhD degree from the Department of Computer Science, University of Toronto in 2014. He assumed his duty as an assistant professor at City University of Hong Kong in 2015. His research interests include bioinformatics, computational biology, evolutionary computation, data mining, machine learning, and interdisciplinary research.

He is merited as the associate editor of BioData Mining in 2016. In addition, he is on the editorial board of Applied Soft Computing since 2016. Remarkably, he has solely edited 2 books published by Springer and CRC Press, attracting 30 peer-reviewed book chapters around the world.

Yi Chang is the Dean of the School of Artificial Intelligence, Jilin University. He has broad research interests on information retrieval, data mining, machine learning, and natural language processing. He has published more than 100 research papers in premium conferences or journals, and he is an associate editor of the *IEEE Transactions on Knowledge and Data Engineering*. He is a senior member of the *IEEE*.

Shengxiang Yang (M'00–SM'14) received the Ph.D. degree from Northeastern University, Shenyang, China in 1999. He is currently a Professor in Computational Intelligence and Director of the Centre for Computational Intelligence, School of Computer Science and Informatics, De Montfort University, Leicester, U.K. He has over 320 publications with an H-index of 55 according to Google Scholar. His current research interests include evolutionary computation, swarm intelligence, artificial neural networks, data mining and data stream mining, and relevant real-world applications. He serves as an Associate Editor/Editorial Board Member of a number of international journals, such as the *IEEE Transactions on Cybernetics*, *IEEE Transactions on Evolutionary Computation*, *Information Sciences*, and *Enterprise Information Systems*.