Multiobjective Patient Stratification Using Evolutionary Multiobjective Optimization

Xiangtao Li and Ka-Chun Wong

***Abstract*—One of the main challenges in modern medicine is to stratify patients for personalized care. Many different clustering methods have been proposed to solve the problem in both quantitative and biologically meaningful manners. However, existing clustering algorithms suffer from numerous restrictions such as experimental noises, high dimensionality, and poor interpretability. To overcome those limitations altogether, we propose and formulate a multiobjective framework based on evolutionary multiobjective optimization to balance the feature relevance and redundancy for patient stratification. To demonstrate the effectiveness of our proposed algorithms, we benchmark our algorithms across 55 synthetic datasets based on a real human transcription regulation network model, 35 real cancer gene expression datasets, and two case studies. Experimental results suggest that the proposed algorithms perform better than the recent state-of-the-arts. In addition, time complexity analysis, convergence analysis, and parameter analysis are conducted to demonstrate the robustness of the proposed methods from different perspectives. Finally, the t-Distributed Stochastic Neighbor Embedding (t-SNE) is applied to project the selected feature subsets onto two or three dimensions to visualize the high-dimensional patient stratification data.**

***Index Terms*—Patient stratification, multiobjective algorithm, clustering.**

I. Introduction

PATIENT stratification or disease subtyping is critical for precision medicine and personalized treatment of complex disease [1], [2]. High-throughput sequencing such as next-generation sequencing (NGS) [3], [4], allows us to sequence massive amounts of various nucleotide and protein data which have revolutionized the study of genomics and molecular biology [5]. Those large-scale studies provide a new and rich data source for addressing ground challenges in cancer research, such as identifying driver genes and stratifying patients into different clinical subtypes [6]. In addition, those subtypes are often connected with various genetic mutations, gene expression profiles, molecular signatures, tissue and organ morphologies as well as different clinical phenotypes [7]. To effectively take care of individual patients, development of novel computational methods for patient stratification leveraging high-throughput molecular data would significantly facilitate personalized healthcare.

Data clustering (e.g., patient stratification) is a task that groups similar objects (e.g. samples) together, which has been widely used for analyzing biological data [8]–[15]; for instance, Miladi *et al.* [16] proposed a method, called RNAscClust, to cluster structured RNAs by considering the structural conservation. Adams *et al.* [17] proposed spatial substitution clustering to identify spatially clustered substitutions at particular branches of phylogenetic trees. Liu *et al.* [18] introduced an Entropy-based Consensus Clustering (ECC) method, which employs an entropy-based utility function to combine different primary partitions to a consensus one. However, most clustering algorithms assume all features to be equally important for clustering. It is one of the reasons why most clustering algorithms may not perform well in the face of high-dimensional molecular data. In reality, various features have diverse effects on clustering [19]. An important feature can help inform clusters while an unimportant feature may not benefit clustering. For the worst case, it may blur the cluster boundaries [20].

As evidenced by the past studies, most features are either noisy or irrelevant and can be removed to reduce the data size for efficient clustering [21], [22]. The task of selecting the "best" feature subset is known as feature selection. Feature selection methods have been widely applied in classification, where the main goal is to select features that can improve the performance [23]. However, feature selection has received comparatively little attention in clustering. Moreover, the existing works [24]–[26] still suffer from numerous drawbacks, such as low optimization efficiency, falling into local optima, and premature convergence. Therefore, future research works are still necessary to develop feature selection methods for clustering.

In order to generate an effective feature subset for clustering, two important aspects should be considered: feature rele-

Manuscript received June 23, 2017; revised September 14, 2017; accepted October 31, 2017. Date of publication November 2, 2017; date of current version August 31, 2018. 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 21200816] and [CityU 11203217]. Also, the work described in this paper was partially supported by a grant from City University of Hong Kong (CityU Project No 7200444/CS), an Amazon Web Service (AWS) Research Grant, and a Microsoft Azure Research Award. This work was supported in part by the National Natural Science Foundation of China under Grant 61603087, in part by the Natural Science Foundation of Jilin Province under Grant 20160101253JC, and in part by the Fundamental Research Funds for Northeast Normal University 2412017FZ026. *(Corresponding author: Ka-Chun Wong.)*

X. T. Li is with the Department of Computer Science and Information Technology, Northeast Normal University, Changchun 130117 China (e-mail: lixt314@nenu.edu.cn).

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

Digital Object Identifier 10.1109/JBHI.2017.2769711

vance and feature redundancy. Traditional methods attempt to maximize relevance and minimize redundancy in different manners. However, unfortunately, those algorithms cannot reveal the tradeoff curves (e.g., Pareto fronts) between feature relevance and feature redundancy. To overcome it, we propose multiobjective feature selection methods based on the state-of-the-art evolutionary multiobjective optimization algorithms for patient stratification. 55 synthetic datasets based on a real human transcription regulation network model, 35 real cancer gene expression datasets, and two case studies are conducted to demonstrate the effectiveness of our proposed algorithms compared with the recent state-of-the-arts.

II. Existing Work

Correlation based measures are frequently utilized for both relevance and redundancy analysis. The most important one is the information theory (such as, mutual information and symmetrical uncertainty) based on the concept of entropy [27], which has been popularly applied in the field of machine learning because of their ability to handle both linear and non-linear relationships.

In information theory, entropy is a measure of the uncertainty or unpredictability in a system. Given a discrete stochastic variable *X*, the entropy of *X* is given by:

$$H(X) = -\sum_{x \in X} p(x)\ \log\ (p(x)); \tag{1}$$

where $p(x) = Pr(X = x)$ is the probability density function of X. For two discrete random variables, X and Y with the joint probability density $p(x, y)$, the joint entropy of a couple of variables can be described as follows:

$$H(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y)\ \log\ (p(x,y)); \tag{2}$$

when some variables are identified, and others are not; the remaining uncertainty is measured by the conditional entropy. Then, the conditional entropy $H(X|Y)$ of X on Y is

$$H(X|Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y)\ \log\ (p(x|y)); \tag{3}$$

The general information that two variables share is determined as the mutual information (MI) between two variables as follows:

$$I(X,Y) = -\sum_{x \in X}\sum_{y \in Y} p(x,y)\ \log\ \frac{p(x,y)}{p(x) \cdot p(y)}; \tag{4}$$

MI is considered to be a good indicator of relevance between two random variables. The MI between X and Y is defined as follows:

$$I(X,Y) = H(X) + H(Y) - H(X,Y); \tag{5}$$

where $I(X,Y)$ is a measure of dependency between variable X and variable Y. It should be normalized between 0 and 1. Therefore, we choose symmetrical uncertainty [28] as a measure of correlation between features and the concept target; it weights features by their symmetrical uncertainty. The features with larger symmetrical uncertainty can get higher weights and vice versa. It can be described as follows:

$$SU(X,Y) = \frac{2I(X,Y)}{H(X) + H(Y)} \tag{6}$$

III. Our Proposed Algorithm

In this section, we present the proposed multiobjective feature selection method based on the state-of-the-art evolutionary algorithms for clustering problems under the context of patient stratification. As we know, high-throughput molecular data often concerns large and high-dimensional data, but most clustering approaches in the literature are sensitive to scalability or high dimensionality or both. Different features affect clusters differently. Some features are important for clusters while others may hinder the clustering task. The most efficient way is to choose a subset of important features. To propose an effective feature selection approach, two important aspects should be considered: feature relevance and feature redundancy. Traditional feature selection algorithms such as Mutual Information based Feature Selection (MIFS) [29] and Normalised Mutual Information Feature Selection (NMIFS) [30] attempt to maximize relevance and minimize redundancy using local search methods. However, those algorithms still have two limitations. Firstly, those algorithms share a general problem: when the number of selected features grows, the redundancy term grows in magnitude. Secondly, a local search strategy cannot guarantee that the best feature subset is decided according to the selection criteria. To address it, we propose a multiobjective feature selection method for clustering the patient stratification data based on the state-of-the-art multiobjective evolutionary algorithms.

1) Objective Functions: In this study, we propose to consider the optimality of feature subset concerning three objectives; our fitness function involves three objective functions. In the filtering phase, those three objective functions include the relevance, the redundancy and the number of features. The first objective function is to maximize the feature relevance. Let X be the set of all features, y be the target label, and Seq be a subset of X, the measure of relevance $f_1(Seq)$ to be maximized is defined as follows:

$$f_1(Seq) = \sum_{x_i \in Seq} SU(x_i, y) \tag{7}$$

where $SU(x_i, y)$ is the symmetric uncertainty between the ith feature and the target label y. The next objective is to minimize the redundancy of the features (i.e., feature regularization), which can be defined as follows:

$$f_2(Seq) = \sum_{x_i, x_j \in Seq, i \neq j} SU(x_i, x_j) \tag{8}$$

where $SU(x_i, x_j)$ is the symmetric uncertainty between the ith feature and the jth feature. The third objective function is the number of selected features, which can be described as follows:

$$f_3(Seq) = |Seq|, |Seq| > 0 \tag{9}$$

Both (7) and (8) evaluate feature subsets as a whole, instead of examining one feature at a time. $f_1(Seq)$ and $f_2(Seq)$ are defined based on the relevance term and redundancy measurement presented in Section II. $f_3(Seq)$ is the number of features within the feature subset of X, which is expected to ensure that the minimization of the number of features is based on the achieved feature subsets with high performance since the objective of feature selection aims to choose a small number of relevant features to achieve similar or even better performance than using all features for model regularization. The multiobjective feature selection algorithm attempts to achieve the following optimization task:

$$min\{-1 \times f_1(Seq)\}, min\{f_2(Seq)\}, min\{f_3(Seq)\} \quad (10)$$

2) Multiobjective Algorithm for Subset Selection: In recent years, many well-known multiobjective evolutionary algorithms have been proven to address different multiobjective optimizations efficiently. In this paper, we extend these algorithms to find the suitable features in the high-dimensional data for clustering the patient stratification data. The overall framework of the proposed algorithm is outlined in Algorithm 1.

As stated in algorithm 1, the Clu-MOEA is constituted of three essential steps. First, a population-based multiobjective evolutionary algorithm is applied to find a set of feature subsets that maximizes the relevance f_1 and minimize the redundancy f_2 and the number of selected features f_3. Then, the multiobjective evolutionary algorithm optimizes those three objective functions. At each generation, the multiobjective evolutionary algorithm calculates those three objective functions for each subset Seq_i and then updates by evolution operators and the environmental selection method in various multiobjective evolutionary algorithms. After that, a Pareto optimal set $\hat{P}$ including all non-dominated individuals on these three objective functions can be found. By optimizing the first and second objective functions, the selected feature subsets not only have high discriminative power, but also have low feature redundancy. By optimizing the objective function f_3, the number of selected features can be limited to balance the tradeoff between relevance and redundancy.

Second, the K-means clustering method is used to cluster the samples with the feature subset Seq_i in Pareto optimal set $\hat{P}$. The K-means clustering method [31] groups items into k groups (where k is the number of pre-chosen groups). Then, each non-dominated individual Seq_i is computed based on the Normalized Mutual Information (NMI_i) and the Normalized Rand Index (R_n^i). Finally, based on the best Normalized Mutual Information value, we apply the t-Distributed Stochastic Neighbor Embedding (t-SNE) [32] to project the feature subset onto two or three dimensions for visualization.

The choice of the multiobjective evolutionary algorithm depends on the algorithmic efficiency, which can balance the diversity and convergence of the population. Two well-known multiobjective evolutionary algorithms including nondominated sorting genetic algorithm II (NSGA-II) [33] and multiobjective evolutionary algorithm based decomposition (MOEA/D) [34] are adopted. In the following section, we introduce each critical component in the algorithm.

Algorithm 1: Pseudocode of Clu-MOEA.

Input: The input dataset X, population size N.
Output: The Normalized Mutual Information (NMI) and the Normalized Rand Index (R_n).
Initial the population P for multiobjective evolutionary algorithms;
while a stopping criteria is not satisfied **do**
  for $i = 1 \rightarrow N$ **do**
    /*calculate the three objective functions for each subset $Seq_i \in P$*/
    compute the -$f_1(Seq_i)$ based on the feature in Seq_i and the target label y for the relevance measure.
    compute the $f_2(Seq_i)$ based on the pair of features in Seq_i for the redundancy measure.
    calculate the number of selected features $f_3(Seq_i) \leftarrow |Seq_i|$
  end for
  Using the update operators and environmental selection in *multiobjective evolutionary algorithm* to make some changes for the next iteration.
  Return the Pareto optimal set $\hat{P}$ including all non-dominated individuals with respect to the three objective functions.
end while
for $i = 1 \rightarrow |\hat{P}|$ **do**
  Using the *K-means clustering method* to cluster the samples with the feature subset Seq_i in $\hat{P}$
  Evaluate the Normalized Mutual Information (NMI_i) and the Normalized Rand Index (R_n^i) for each non-dominated individual Seq_i
end for
Find the best Normalized Mutual Information value $NMI = max_{i \in |\hat{P}|}(NMI_i)$ and the best Normalized Rand Index R_n.
Apply the *t-Distributed Stochastic Neighbor Embedding (t-SNE)* to project the best feature subset obtained Normalized Mutual Information value or Normalized Rand Index onto two or three dimensions to visualization of the patient stratification data.

1) *NSGA-II*: NSGA-II is a well-known multi-objective meta-heuristic algorithm, which utilizes not just an elite-preserving strategy but also a specific diversity-preserving mechanism. This algorithm applies a population of individuals and obtains its Pareto front. Two major operators of NSGA-II are non-dominated sorting and crowding distance procedures. In each generation, the offspring is created by using the parent population with simulated binary crossover and polynomial mutation operators. Instead of finding the individuals only in the non-dominated solutions, these two populations are first combined. Then, a fast non-dominated sorting method is used to classify the population. After that, the new population is filled with the solutions of different non-dominated fronts, one at a time. To preserve pop-

ulation diversity, the crowding distance method is used to choose the suitable individual into the next generation of the algorithm. The algorithm is stopped after reaching the stopping criteria (e.g., the maximal number of generations).

2) *MOEA/D*: MOEA/D is a new MOEA framework, which has been popularly used for solving many multiobjective optimization problems. It decomposes a multiobjective optimization problem into multiple scalar optimization subproblems based on different weight vectors. Then, the algorithm can handle these subproblems simultaneously by evolving a population of solutions. At the beginning of the algorithm, a neighborhood of weight vector λ is represented as a set of its several closest weight vectors. The neighborhood of each subproblem consists of all subproblems with the weight vectors. Each subproblem has its best solution found so far in the population. Meanwhile, simulated binary crossover and polynomial mutation operators can be used to generate the new individual. Then, the new individual is compared with the original solution with the same weight vector and its neighboring subproblems. When the algorithm satisfies a given stopping criteria, the algorithm outputs the Pareto optimal set.
3) *t-Distributed Stochastic neighbor embedding methodology (t-SNE)*: The t-SNE algorithm is a machine learning algorithm for dimensionality reduction, which is particularly well suited for the visualization of high-dimensional data. This algorithm is executed by applying the Barnes-Hut approximations, which calculates the similarities of the high-dimensional data points assuming Student-t distribution as a distance measure and projects the data onto lower dimensions while preserving the similarities. It converts similarities between data samples to joint probabilities and minimize the Kullback-Leibler divergence.

A. Time Complexity analysis

In this section, we focus on the time complexities of proposed methods. In the Clu-MOEA algorithm, it includes three essential parts: multiobjective evolutionary algorithm, K-mean clustering algorithm, and t-Distributed stochastic neighbor embedding methodology (t-SNE).

For the multiobjective evolutionary algorithm, two well-known multiobjective evolutionary algorithms including nondominated sorting genetic algorithm II (NSGA-II) and multiobjective evolutionary algorithm based decomposition (MOEA/D) have been taken. For NSGA-II, in order to generate offsprings, simulated binary crossover and polynomial mutation operators are used, costing $O(N \cdot D)$. Since each individual dominates (N–1) other individuals at maximum and each domination check requires at most M comparisons, the overall time complexity is $O(M \cdot N^2)$, where N is the population size and the M is the number of objectives. To choose the suitable individuals into the next generation, the crowding distance method is used with the time complexity $O(M(2N)\ \log\ (2N))$. For each iteration of NSGA-II algorithm, the overall time complexity is $O(M \cdot N^2 + N \cdot D)$, which is dominated by the nondominated sorting of the algorithm. For the MOEA/D, the algorithm randomly picks two solutions for genetic operators and find the reference point, costing $O(M)$. For the update of neighboring solutions, it needs $O(MT)$ underlying operations since its major costs are computed with T solutions. Simulated binary crossover and polynomial mutation operators are applied to generate the offsprings, which costs $O(N \cdot D)$. Therefore, the computational complexity of MOEA/D is $O(M \cdot N \cdot T + N \cdot D)$ since it has N individuals with D dimensions, where T is the number of the weight vectors in the neighborhood of each weight vector. If the maximal iteration is G, the overall time complexity of proposed multiobjective algorithm based on NSGA-II is $O((M \cdot N^2 + N \cdot D) \cdot G)$ and the overall time complexity of proposed multiobjective algorithm based on MOEA/D is $O((M \cdot N \cdot T + N \cdot D) \cdot G)$.

For the K-mean clustering problem, if the number of clusters k, the number of features D, and the number of samples S are fixed, the problem can be accurately solved in time $O(k \cdot D \cdot S \cdot I)$ with a fixed number I of iterations. For t-Distributed Stochastic neighbor embedding methodology (t-SNE), the algorithm reduces the high-dimensional data into two-dimensional spaces using Barnes-Hut Stochastic Neighbor Embedding algorithm with the time complexity of $O(N \cdot \log\ (N))$.

The overall time complexity of CNSGA-II based on NSGA-II (CNSGA-II) is $O((M \cdot N^2 + N \cdot D) \cdot G + k \cdot D \cdot S \cdot I + N \cdot \log\ (N))$. The overall time complexity of CMOEA/D based on MOEA/D (CMOEA/D) is $O((M \cdot N \cdot T + N \cdot D) \cdot G + k \cdot D \cdot S \cdot I + N \cdot \log\ (N))$.

IV. Experiments

A. Data Sources

We simulate 55 synthetic gene expression datasets from the published dynamical gene regulation model [18]. All datasets are generated as follows:

$$F_i^{mRNA}(x,y) = \frac{dx_i}{dt} = m_i \cdot f_i(y) - \lambda_i^{mRNA} \cdot x_i,$$
$$F_i^{Prot}(x,y) = \frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{Prot} \cdot y_i,$$
$$i = 1, \cdot, n \tag{11}$$

where m_i is the maximum transcription rate and r_i is the translation rate, λ_i^{mRNA} and λ_i^{Prot} are the mRNA and protein degradation rates. $f_i(\cdot)$ is the relative activation of gene. For the 55 synthetic gene expression datasets, they are provided based on a real human transcription regulation network. Each dataset contains 200 samples, which are classified into 4 clusters. All parameter settings follow [18]. The number of knock-out genes is varied from 100 to 500. The noise level is varied from 0 to 0.5. Each knock-out genes includes 11 instances for different noise levels.

In addition, we also adopted the gene expression datasets with label information from the benchmark study in [18]. 35

TABLE I
KEY CHARACTERISTICS OF 35 BENCHMARK DATASETS FOR CLUSTER VALIDITY

No.	Datatset	Tissue	Subject	Gene	Class	No.	Datatset	Tissue	Subject	Gene	Class
1	Alizadeh-2000-v1	Blood	42	1095	2	19	Lapointe-2004-v2	Prostate	110	2496	4
2	Alizadeh-2000-v2	Blood	62	2093	3	20	Liang-2005	Brain	37	1411	3
3	Alizadeh-2000-v3	Blood	62	2093	4	21	Nutt-2003-v1	Brain	50	1377	4
4	Armstrong-2002-v1	Blood	72	1081	2	22	Nutt-2003-v2	Brain	28	1070	2
5	Armstrong-2002-v2	Blood	72	2194	3	23	Nutt-2003-v3	Brain	22	1152	2
6	Bhattacharjee-2001	Lung	203	1543	5	24	Pomeroy-2002-v1	Brain	34	857	2
7	Bitter-2000	Skin	38	2201	2	25	Pomeroy-2002-v2	Brain	42	1379	5
8	Bredel-2005	Brain	50	1739	3	26	Ramaswamy-2001	Multi-tissue	190	1363	14
9	Chen-2002	Liver	179	85	2	27	Risinger-2003	Endometrium	42	1771	4
10	Chowdary-2006	Multi-tissue	104	182	2	28	Shipp-2002	Blood	77	798	2
11	Dyrskjot-2003	Bladder	40	1203	3	29	Singh-2002	Prostate	102	339	2
12	Garber-2001	Lung	66	4553	4	30	Su-2001	Multi-tissue	174	1571	10
13	Golub-1999-v1	Bone marrow	72	1868	2	31	Tomlins-2006-v1	Prostate	92	1288	4
14	Golub-1999-v2	Bone marrow	72	1868	3	32	Tomlins-2006-v2	Prostate	104	2315	5
15	Gordon-2002	Lung	181	1626	2	33	West-2001	Breast	49	1198	2
16	Khan-2001	Multi-tissue	83	1069	4	34	Yeoh-2002-v1	Bone marrow	248	2526	2
17	Laiho-2007	Colon	37	2202	2	35	Yeoh-2002-v2	Bone marrow	248	2526	6
18	Lapointe-2004-v1	Prostate	69	1625	3						

benchmark cancer gene expression datasets [35] presented in Table I are adopted to benchmark the clustering ability of Clu-MOEA, comparing with different clustering methods. Among the 35 benchmark datasets, the number of samples is varied from 22 to 248; the number of features is varied from 85 to 4553; and the number of clusters is varied from 2 to 14. Some datasets are from the same data source such as the Alizadeh-2000-v2 and Alizadeh-2000-v3 use the same data. The Yeoh-2002-v1 and Yeoh-2002-v2, Golub-1999-v1 and Golub-1999-v2 are originated from the same study but have different clusters.

Finally, two case studies including three breast cancer gene expression datasets are adopted for detailed evaluations of our proposed algorithms.

B. Parameter Setting

The comparisons are conducted on three different kinds of datasets: the first is 55 synthetic data, and the second is 35 cancer gene expression benchmark datasets and the last is two case studies including three high-dimensional cancer gene expression datasets. In this paper, we use multiobjective feature section method Clu-MOEA to patient stratification problems. For the Clu-MOEA, two well-known multiobjective algorithms including NSGA-II and MOEA/D are adopted in this paper. For both algorithms, the population size (N) is 20. In other aspects, to make the experiments accurate, both algorithms have been run for 30 independent times on each dataset. Then, we calculate the average result of 30 independent runs. For the stopping criteria, the number of fitness evaluations in each generation depends on the operators used and the population update model. Different operators can lead to very different numbers of function evaluations per generation. That is the reason why it is obvious to use the number of function evaluations as the stop criterion instead of the number of generations or CPU times. Therefore, in this paper, the number of function evaluations is the stopping (i.e., termination) criteria. We set 1000 objective function evaluation for each dataset. For NSGA-II, the mutation rate is 0.1. For the MOEA/D, the number of the weight vectors in the neighborhood of each weight vector T is 4, and the mutation rate is also 0.1.

C. Other Related Methods from Literature

To demonstrate the effectiveness and efficiency of Clu-MOEA, we compare our proposed algorithm with six different clustering methods; namely, K-mean clustering method (KM) [31], Agglomerative hierarchical clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL) [36], Link-based Cluster Ensemble (LCE) [37], Entropy based Consensus Clustering (ECC) [18]. Meanwhile, we also compare our proposed algorithm with multiobjective particle swarm optimization algorithm [38]. For all datasets, we have benchmarked each method as shown in Figs. 1–3 and Supplementary Table S1 and S2. From the experimental results, it is noted that CMOEA/D can provide the best solutions among these eight algorithms. To assess its performance rigorously, we use the Paired Wilcoxon signed rank test [39], [40] to compare our algorithms with other algorithms to statistically ascertain whether the experiment results of our algorithms are statistically better than other algorithms. The pair Wilcoxon signed-rank test is a non-parametric statistical hypothesis test, which can be used as an alternative to the pair Wilcoxon signed rank test when the results cannot be assumed to be normally distributed. In the pair Wilcoxon signed rank test, the null hypothesis represents no significant improvement compared with other algorithms, and the alternative hypothesis represents that our algorithm is statistically different from other methods.

D. Evaluation metrics

Two important metrics are computed in our experiments, which are Normalized Mutual Information (NMI) [41] and Normalized Rand Index (R_n) [42]. All these metrics are applied to measure the similarities between the cluster labels and

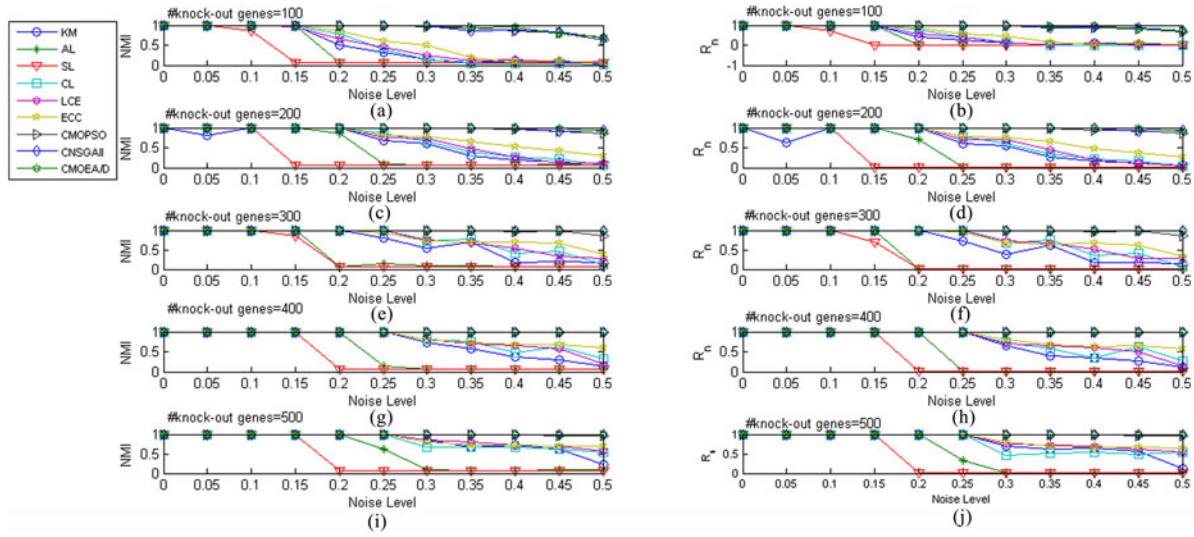

Fig. 1. Performance of different clustering algorithms on the 55 synthetic datasets (based on a real human transcriptional regulation network of 2723 genes). CMOEA/D has substantial advantages over other methods on the datasets. The number of knock-out genes varied from 100 to 500. The noise level varied from 0 to 0.5. Each knock-out genes includes 11 instances for different noise levels. (a) and (b) are the results of NMI and R_n with 100 knock-out genes. (c) and (d) are the results of NMI and R_n with 200 knock-out genes. (e) and (f) are the results of NMI and R_n with 300 knock-out genes. (g) and (h) are the results of NMI and R_n with 400 knock-out genes. (i) and (j) are the results of NMI and R_n with 500 knock-out genes.

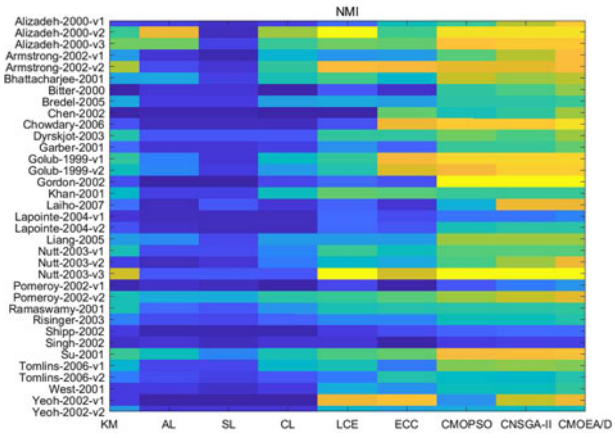

Fig. 2. The performance of CMOEA/D on 35 benchmark cancer gene expression data sets on NMI. The horizontal axis denotes different clustering algorithms while the vertical axis denotes the 35 cancer gene expression data.

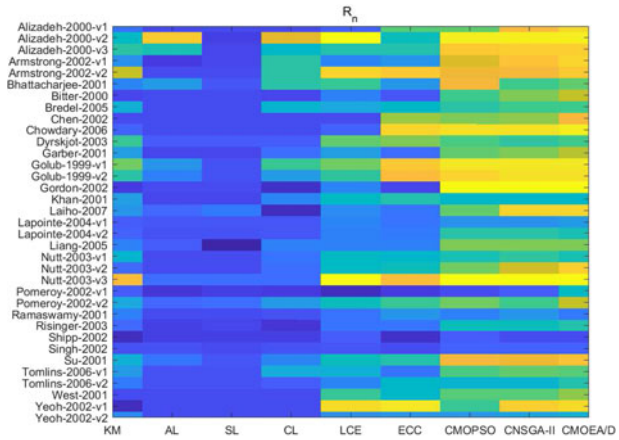

Fig. 3. The performance of CMOEA/D on 35 benchmark cancer gene expression data sets on R_n. The horizontal axis denotes different clustering algorithms while the vertical axis denotes the 35 cancer gene expression data.

the truth labels. The results with higher values represent better partitions than other results.

Normalized Mutual Information is a reliable criterion in the information-theoretic metric of the agreement between the real and predicted results with normalization to assure NMI ranges from 0 and 1. It can be defined as follows:

$$NMI = \frac{\sum_{i,j} n_{i,j} \log \frac{n \cdot n_{i,j}}{n_{i+} \cdot n_{+j}}}{\sqrt{(\sum_i n_{i+} \log \frac{n_{i+}}{n})(\sum_j n_{j+} \log \frac{n_{+j}}{n})}} \quad (12)$$

Normalized Rand Index is a measure of agreement between partitions and the target label, which varies between 0 and 1 according to the expectation of random partitions:

$$R_n = \frac{\sum_{i,j} \binom{n_{i,j}}{2} - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{\sum_i \binom{n_{i+}}{2}/2 + \sum_j \binom{n_{+j}}{2}/2 - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}} \quad (13)$$

It is remarked that NMI and R_n are positive measurements. Although R_n is normalized, it can still be negative; it means that the results are even worse than the random cases.

E. Synthetic data

In this section, we compare our proposed algorithms with other clustering methods on the 55 synthetic datasets based on a real human transcriptional regulation network of 2723 genes. When we perform feature selection processes of our

proposed algorithm, the numbers of features to be selected by both algorithms are used to balance the feature relevance and feature redundancy. The results of NMI and R_n are given in Fig. 1 over the 30 independent runs. We can observe that our proposed algorithms CNSGA-II and CMOEA/D have substantial advantages over other methods on the datasets. Comparing these two algorithms, the results of CMOEA/D can provide better solutions than that of CNSGA-II on these synthetic datasets.

F. Benchmark on cancer gene expression data

In the previous section, we have employed CNSGA-II and CMOEA/D to stratify synthetic disease data. In this section, we use these two methods to cluster the 35 cancer gene expression datasets. The experimental results are summarized in Supplementary Table S1 and S2. For the evaluation metrics NMI, Supplementary Table S1 shows the average results of the 30 runs for each method. Comparing CNSGA-II and CMOEA/D, we can find that CNSGA-II is inferior to, equal to and superior to COMEA/D in 23, 2, and 10 cancer gene expression data respectively. The "+" denotes our algorithm CMOEA/D is better than other algorithms. And the "-" and "≈" denote our algorithm is inferior to, equal to other algorithms. Out of the eight algorithms for comparison, the CMOEA/D is statistically better than other algorithms labeled "+". Meanwhile, from the results, it can be seen that CMOEA/D and CNSGA-II are superior to all other traditional clustering methods. For Agglomerative hierarchical clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL), our proposed algorithms are statistically significantly better as compared to all these algorithms except our proposed methods. For the Complete-Linkage, it can produce better solutions than other algorithms. Comparing Link-based Cluster Ensemble, our algorithm CMOEA/D can be inferior to, equal to and superior to this algorithm on 1, 3, and 31 datasets, respectively. For the entropy-based consensus clustering (ECC), CMOEA/D can perform better on 33 cancer gene expression datasets. For CMOPSO, CMOEA/D can give the better solutions on 28 cancer gene expression datasets. Meanwhile, our proposed algorithm produces promising results on several datasets by a larger margin, such as Armstrong-2002-v2, Garber-2001, and Laiho-2007. Although LCE and ECC can lead to acceptable results on several datasets, they are weaker in the robustness, such as LCE can perform on Alizadeh-2000-v2, but produce the worst solutions in Pomeroy-2002-v1. It is noted that there are specific datasets (Gordon-2002, Pomeroy-2002-v1, Risinger-2003) for which the traditional clustering methods yields very poor performance but our proposed algorithms can produce far better solutions, most likely due to the presence of irrelevant or noisy features. Fig. 2 shows the clustering performance of different algorithms measured by NMI. Therefore, it can conclude that CMOEA/D is better than other algorithms which indicate that CMOEA/D has the greater robustness than other methods on the majority of cancer gene expression data.

In other aspects, the experimental results of R_n are summarized in Supplementary Table S2 which tabulates the mean values of 30 runs for each method. As observed from Supplementary Table S2, several observations can be made: (1) our proposed CNSGA-II and CMOEA/D obtain better solutions than other clustering algorithms while CMOEA/D performs better than the CNSGA-II algorithm in 20 cancer gene expression datasets. Meanwhile, CNSGAI-II algorithm performs better than CMOEA/D on 7 cancer gene expression datasets. (2) Agglomerative hierarchical clustering with single-Linkage (SL) obtains the worst performance among all these algorithms while CMOEA/D can provide better solutions on all 35 cancer gene expression dataset. Meanwhile, agglomerative hierarchical clustering with complete-Linkage (CL) provides better solutions than Single-Linkage (SL) and Average-Linkage (AL). (3) Entropy-based Consensus Clustering (ECC) can give better solutions than other traditional algorithms while CMOEA/D is inferior to, equal to and superior to ECC in 0, 5, and 30 cancer gene expression datasets, respectively. (4) CMOPSO can provide better solutions than Entropy-based Consensus Clustering while CMOEA/D is inferior to, equal to and superior to 7, 4, 24 cancer gene expression datasets. (5) For some larger scale cancer gene expression datasets such as Bitter-2000 and Laiho-2007, CMOEA/D can give the best performance while other algorithms perform unsatisfactorily. (5) Fig. 3 shows the clustering performance of different algorithms measured by R_n, which demonstrates our algorithm CNSGA-II and CMOEA/D can perform better than all other algorithms.

Based on the above analysis, we can conclude that our proposed algorithms own significant advantages in patient stratification. It also demonstrates that our algorithms can alleviate the detriment effect of irrelevant and noisy features via the multiobjective approach.

G. Convergence analysis

From the Supplementary Tables S1 and S2, we can observe that CMOEA/D can achieve the best performance on the 35 cancer gene expression datasets. To investigate the full performance spectrum of the CMOEA/D algorithm, a convergence analysis based on the number of objective function evaluations (FEs) is conducted on these 35 cancer gene expression datasets. The average values of all these 35 cancer gene expression datasets are obtained against different numbers of objective function evaluations (FEs). The experimental results are depicted in Fig. 4. As can be seen in Fig. 4, we can conclude that the performance of CMOEA/D can be enhanced along with growing number of objective function evaluations, demonstrating its potential with the increasing computing power in the future.

H. Parameter analysis

1) Stability of population size: Population sizing has been one of the major topics to be considered in the evolutionary computation. The task of selecting an appropriate population size for solving particular classes of problems has been known to be a challenge in the evolutionary computing community. Researchers usually claim that using a small population size may lead the algorithm to poor solutions and that increasing the population size will improve the diversity of possible changes and promote the exploration of the search space. Due to the significant influence of population size on the solution quality

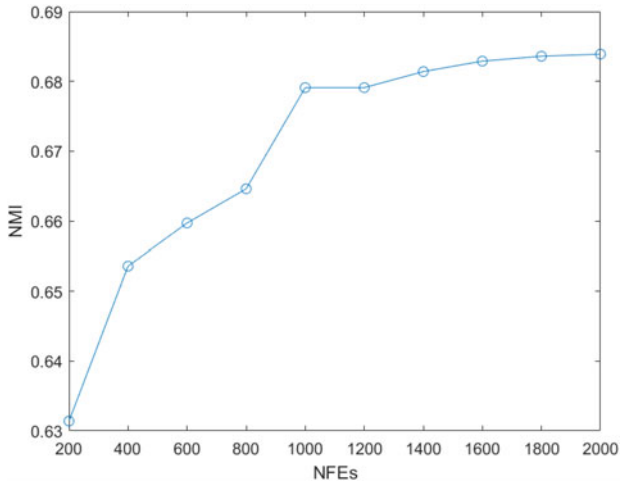

Fig. 4. Convergence behavior of CMOEA/D. The horizontal axis denotes the number of objective function evaluations while the vertical axis denotes the normalized mutual information (NMI).

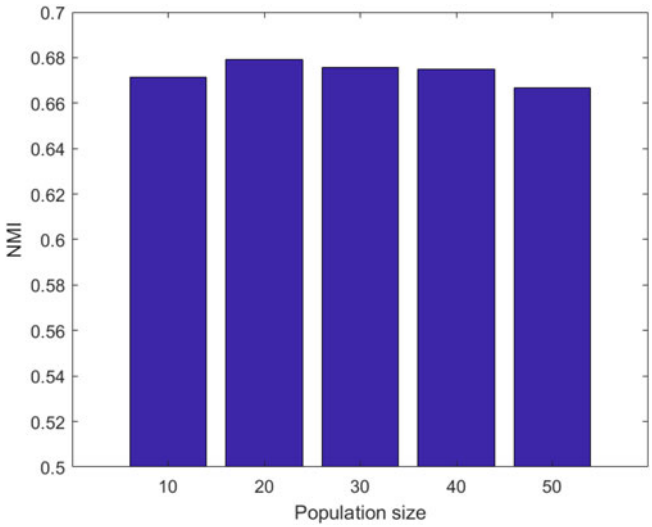

Fig. 5. Parameter analysis on the population size of CMOEA/D under 1000 objective function evaluations. The horizontal axis denotes different population sizes while the vertical axis denotes the normalized mutual information (NMI).

and search time, parameter analysis experiments should be conducted. In this section, we use the evaluation metric NMI as the criterion to compare different settings.

To reveal empirical insights into the impact of the population size on the search performance of CMOEA/D, the population size has been varied from 10 to 50 on 35 cancer gene expression datasets under the same number of objective function evaluations 1000. The results of NMI are summarized in Fig. 5. From Fig. 5, we can observe that CMOEA/D with the population size 20 provides the best solutions.

2) Sensitivity of T in CMOEA/D: To study the sensitivity to the parameter T in CMOEA/D on 35 cancer gene expression datasets, we have tested different settings of T in the CMOEA/D algorithm in Fig. 6 concerning NMI. As clearly shown in Fig. 6, CMOEA/D performs very well with $T = 4$ on those

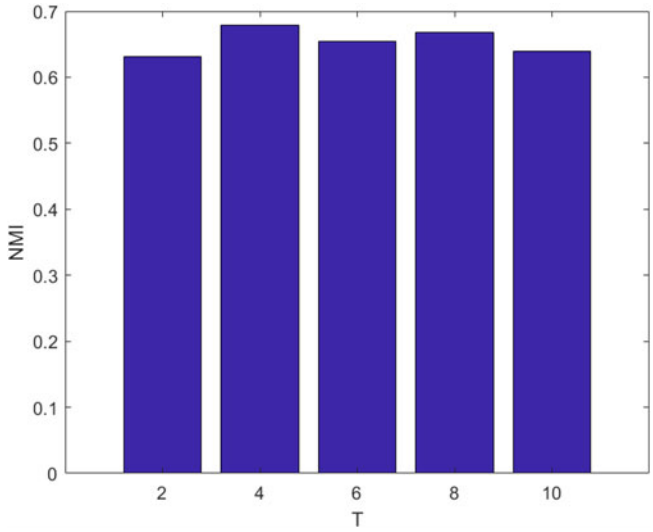

Fig. 6. The average NMI value versus the value of T in CMOEA/D for 35 cancer gene expression datasets. The horizontal axis denotes different neighborhood sizes while the vertical axis denotes the normalized mutual information (NMI).

datasets. Fig. 6 also reveals that CMOEA/D does not work well on datasets when T is very small, resulting in poor performance of the population-based search in CMOEA/D.

V. Application

A. Unsupervised Feature Selection

In this experiment, we use the entropy measure $H(X)$ in the (1) instead of the objective function $f_1(seq)$ to design an unsupervised learning algorithm for patient stratification dataset. In the entropy, $H(X) = 0$ denotes that $Pr(X = x) = 1$, where x is the sole value of X occurring in the dataset. In this case, it is reasonable to believe that X has a stable distribution, and will be likely to take the value of x. In contrary, a higher $H(X)$ corresponds to a more random distribution than the previous case. In order to show the difference between the Unsupervised-CMOEA/D (Un-CMOEA/D) and CMOEA/D, 35 real cancer gene expression datasets are applied. The experiment results are summarized in Supplementary Table S3-S4 and Fig. 7. The results in these tables and figure show that Un-CMOEA/D and CMOEA/D have similar performance results for cancer gene expression dataset. For NMI, CMOEA/D can be inferior to, equal to, and superior to Un-CMOEA/D on 10, 9, and 13 datasets respectively. For R_n, CMOEA/D can provide better solutions on 15 cancer gene expression dataset while Un-CMOEA/D can obtain 11 better cancer gene expression dataset. Meanwhile, compared with Supplementary Table S1 and S2 including KM, AL, SL, CL, LCE, ECC, CMOPSO, and CNSGA-II, Un-CMOEA/D can still perform significantly better than the other algorithms on most cancer gene expression dataset, which is similar to CMOEA/D.

B. Visualization of Patient Stratification

Visualization plays an essential role in the field of machine learning, especially when the quantity of data is large. Based on

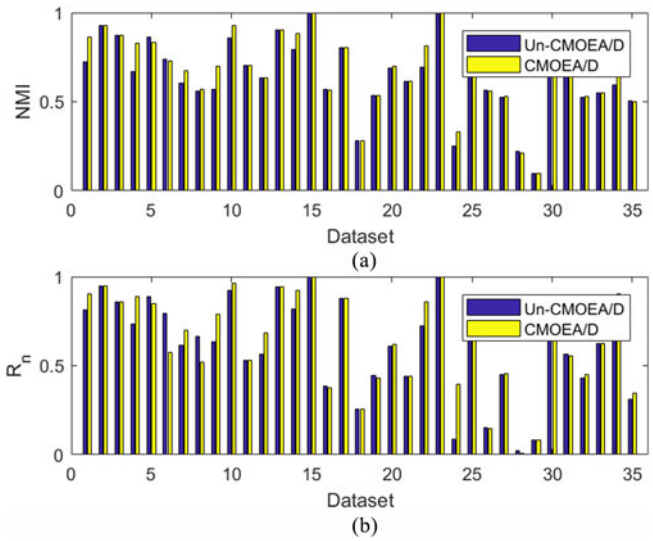

Fig. 7. The average NMI values and R_n values of Un-CMOEA/D and CMOEA/D for 35 cancer gene expression datasets. The horizontal axis denotes different datasets while the vertical axis denotes the normalized mutual information (NMI) or R_n. (a) denotes the average NMI values for 35 cancer gene expression datasets; (b) denotes the average R_n values for 35 cancer gene expression datasets.

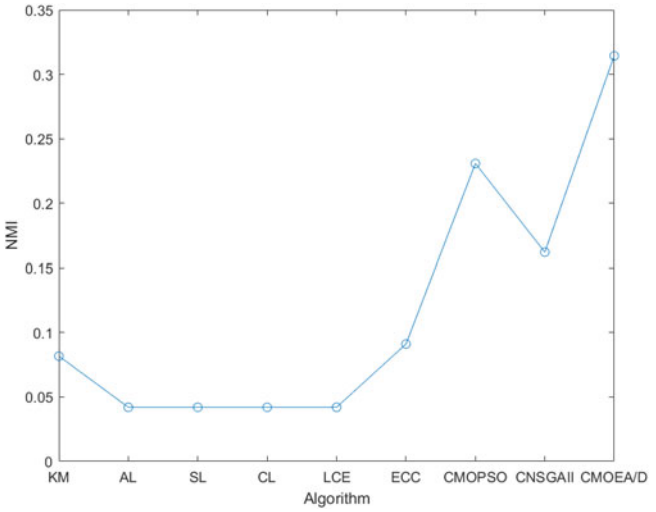

Fig. 8. Performance of different clustering algorithms on Breast Cancer dataset by NMI. The horizontal axis denotes different clustering algorithms while the vertical axis denotes the normalized mutual information (NMI).

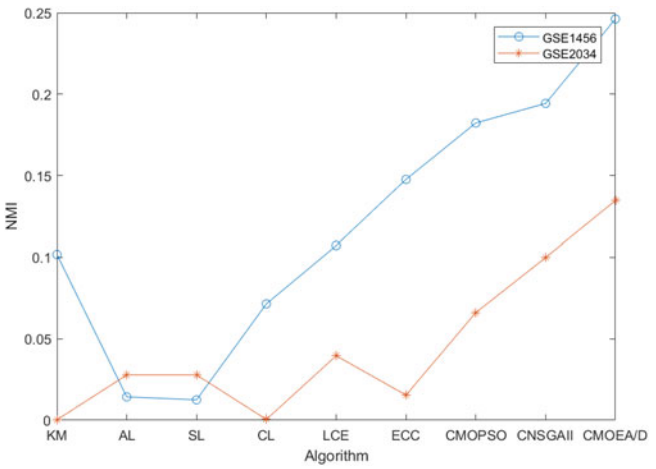

Fig. 9. Performance of different clustering algorithms on GSE1456 and GSE2034 by NMI. The horizontal axis denotes different clustering algorithms while the vertical axis denotes the normalized mutual information (NMI).

the 35 cancer gene expression datasets, t-Distributed Stochastic Neighbor Embedding (t-SNE) is used to visualize all datasets. To demonstrate the effectiveness of our proposed algorithm, the principal component analysis (PCA) is also used for dimension reduction. The comparisons of 2D visualization are shown in Supplementary Figure S1 and Supplementary Figure S2. The axes are in arbitrary units. Each point represents a sample in the cancer gene expression datasets. None of the algorithms used the true labels, and the true label information was added in the form of distinct colors to validate the results. As depicted in Supplementary Figure S1 and Supplementary Figure S2, we can find that t-SNE can successfully divide the patient samples into different subgroups consistent with the original dataset.

C. Case studies

Two case studies are conducted to demonstrate the unique performance of our proposed algorithms. The first case study is conducted on the Breast Cancer dataset, which consists of 97 samples and 24482 genes [43]. Data are divided into two groups; 19 control samples (12 cases are related to relapse samples, and 7 cases are related to nonrelapse samples) used in the test process, and 78 cancer samples (34 cases are related to relapse samples, and 44 cases are related to nonrelapse samples). Then, we employ our proposed algorithm to select the suitable features for patient stratification. The feature relevance and redundancy for each subset are calculated. 164 features were found to be significantly associated with the disease outcomes. In comparison to other methods, our proposed algorithm is associated with a smaller number of currently known cancer-related genes for clustering. Since disease status is not merely related to the number of these cancer-related genes, we obtain the clustering performance by NMI. The experimental results are shown in Fig. 8. From Fig. 8, the proposed algorithm can obtain the NMI value larger than 0.3, while most clustering method including K-mean clustering method (KM), Agglomerative hierarchical clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL), Link-based Cluster Ensemble (LCE), Entropy based Consensus Clustering (ECC), are less than 0.1. It can conclude that our algorithm can perform the best while other algorithms seem to fail in clustering this dataset.

In addition, we collect other breast cancer datasets from Gene Expression Omnibus including GSE1456, and GSE2034, which contains thousands of genes (the number of gene features $>$ 10000) [44]. This dataset addresses the problem of clustering metastatic relapses in breast cancer in different cohorts and is obtained with the Affymetrix HG-U133A technology. The results of all clustering algorithms are listed in Fig. 9. From

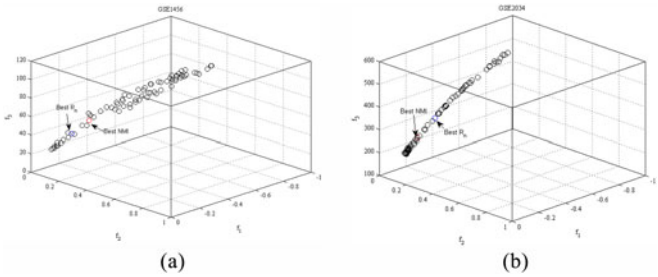

Fig. 10. The Pareto front where each point is associated with a feature subset for GSE1456 and GSE2034. The best NMI is marked by using the "red". Each axis represents one objective function to be optimized under our proposed multiobjective patient stratification framework. Details can be found in the main text. (a) is the is associated with a feature subset for GSE1456 and (b) is associated with a feature subset for GSE2034

the experimental results, our proposed algorithm can provide better solutions than other algorithms, especially for the Agglomerative hierarchical clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL). To further demonstrate the effectiveness of feature selection, we draw the Pareto front where each point is associated with a feature subset in Fig. 10. Then, we marked the best *NMI* value using the "Red" colour. From the results of GSE1456, the best NMI includes 42 features. The top features include 'CX3CR1' [45], 'E2F1' [46], 'HMMR' [47], and 'CDKN1C' [48] which are highly related to the breast cancer; for instance, CX3CR1 protein is detected in human tissue microarrays of normal and malignant mammary glands; breast cancer cells expressing high levels of CX3XR1 have a higher propensity to spread to the skeleton [45]. E2F1 expression is related with the poor survival of lymph node-positive breast cancer patients treated with fluorouracil, doxorubicin, and cyclophosphamide [46]. For the best R_n of GSE1456, the top features include 'E2F1' [46], 'HMMR' [47], 'CDKN1C' [48], and 'GPR56' [49]. The first three features are the same with the best *NMI*. The last feature 'GPR56' is different with the best *NMI*. This feature plays varying roles in endogenous cancer progression, which suppressed cancer progression in the TRAMP model on a mixed genetic background [49]. From the results of GSE2034, the best *NMI* includes 123 features. The top features include 'RACGAP1' [50], and 'GNG7' [51]. The 'RACGAP1' is a GTPase activating protein (GAP) which co-localizes with the mitotic spindle in the metaphase and is essential for cytokinesis during the normal cell cycle in different breast cancer [50]. Large G protein gamma 7 (GNG7) was downregulated in cancer patients, which determined whether its expression was altered at the genomic level, mRNA level, epigenetic level or the microRNA level [51]. For the best R_n, the top features include 'CEP55' [52], 'GNG7' [51] and 'RAB22A' [53]. The CEP55 is a microtubule-bundling protein, which has been recognized recently in several human cancers [52]. The GNG7 is the same feature with the best *NMI*. The RAB22A is an independent adverse prognostic biomarker for breast cancer patients association with the expression of other genes [53]. From the results of *NMI* and R_n of GSE2034, we can conclude that GNG7 has an important role in clustering the GSE2034.

VI. Conclusion

In summary, we demonstrate that our proposed multiobjective framework owns unique and competitive edges over existing clustering algorithms for patient stratification. In particular, our proposed population-based algorithms can reveal a multitude of relevant feature subsets for patient stratification under multiple objectives by removing irrelevant, redundant, and noisy features simultaneously. Moreover, our algorithms have better performance than the others on the high-dimensional cancer gene expression data. In addition, we have adopted the t-Distributed stochastic neighbor embedding methodology (t-SNE) to transform the feature subset onto low dimensions for patient stratification visualization. To demonstrate the robustness, we have conducted the convergence analysis and parameter analysis for the proposed algorithms. Lastly, the proposed algorithms have been extensively benchmarked on 55 synthetic datasets based on a real human transcription regulation network, 35 real cancer gene expression datasets, and two case studies, demonstrating its unique and complementary performance in patient stratification under the proposed multiobjective framework.

Acknowledgment

The authors would like to thank Dr. Hongfu Liu for making their data publicly available. The authors also would like to thank the three anonymous reviewers for their constructive comments which have improved the study in numerous ways.

References

[1] M. Uhlen, B. Hallstrom, C. Lindskog, A. Mardinoglu, F. Ponten, and J. Nielsen, "Transcriptomics resources of human tissues and organs." *Molecular Syst. Biol.*, vol. 12, no. 4, 2016, Art. no. 862.
[2] Q. Zhu *et al.*, "Targeted exploration and analysis of large cross-platform human transcriptomic compendia." *Nature Methods*, vol. 12, no. 3, pp. 211–214, 2015.
[3] S. C. Schuster, "Next-generation sequencing transforms today's biology." *Nature Methods*, vol. 5, no. 1, pp. 16–18, 2008.
[4] E. R. Mardis, "The impact of next-generation sequencing technology on genetics," *Trends Genetics*, vol. 24, no. 3, pp. 133–141, 2008.
[5] J. W. Davey, P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, and M. L. Blaxter, "Genome-wide genetic marker discovery and genotyping using next-generation sequencing," *Nature Rev. Genetics*, vol. 12, no. 7, pp. 499–510, 2011.
[6] H. B. Frieboes, X. Zheng, C. H. Sun, B. Tromberg, R. Gatenby, and V. Cristini, "An integrated computational/experimental model of tumor invasion," *Cancer Res.*, vol. 66, no. 3, pp. 1597–1604, 2006.
[7] E. Bresin *et al.*, "Combined complement gene mutations in atypical hemolytic uremic syndrome influence clinical phenotype," *J. Amer. Soc. Nephrol.*, vol. 24, no. 3, pp. 475–486, 2013.
[8] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering." *Bioinformatics*, vol. 31, no. 22, pp. 3718–3720, 2015.
[9] Sehi L'Yi *et al.*, "XCluSim: A visual analytics tool for interactively comparing multiple clustering results of bioinformatics data," *BMC bioinformatics*, vol. 16, no. 11, 2015, Art. no. S5.
[10] L. Jiang, Y. Dong, N. Chen, and T. Chen, "DACE: A scalable DP-means algorithm for clustering extremely large sequence data," *Bioinformatics*, vol. 33, no. 6, pp. 834-842, 2016.
[11] D. H. Milone, G. Stegmayer, M. Lpez, L. Kamenetzky, and F. Carrari, "Improving clustering with metabolic pathway data," *BMC bioinformatics*, vol. 15, no. 1, 2014, Art. no. 101.
[12] C. Xu, and Z. Su, "Identification of cell types from single-cell transcriptomes using a novel clustering method," *Bioinformatics*, vol. 31, no. 12, pp. 1974–1980, 2015.

[13] D. Jaeger, J. Barth, A. Niehues, and C. Fufezan, "pyGCluster, a novel hierarchical clustering approach," *Bioinformatics*, vol. 30, no. 6, pp. 896–898, 2014.

[14] 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, 2014, Art. no. 37.

[15] P. A. Jaskowiak, R. J. Campello, and I. G. Costa, "On the selection of appropriate distances for gene expression data clustering," *BMC Bioinformat.*, vol. 15, no. 2, pp. S2, 2014.

[16] M. Miladi *et al.*, "RNAscClust: clustering RNA sequences using structure conservation and graph based motifs," *Bioinformatics*, vol. 33, no. 14, pp. 2089–2096, 2017.

[17] J. Adams, M. J. Mansfield, D. J. Richard, and A. C. Doxey, "Lineage-specific mutational clustering in protein structures predicts evolutionary shifts in function," *Bioinformatics*, vol. 33, no. 9, pp. 1338–1345, 2017.

[18] 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.

[19] G. Chandrashekar and F. Sahin, "A survey on feature selection methods," *Comput. Elect. Eng.*, vol. 40, no. 1, pp. 16–28, 2014.

[20] Q. Song, J. Ni, and G. Wang, "A fast clustering-based feature subset selection algorithm for high-dimensional data," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 1, pp. 1–14, Jan. 2013.

[21] Z. Li, J. Liu, Y. Yang, X. Zhou, and H. Lu, "Clustering-guided sparse structural learning for unsupervised feature selection," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 9, pp. 2138–2150, Sep. 2014.

[22] V. Boln-Canedo, N. Snchez-Marono, A. Alonso-Betanzos, J. M. Bentez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Inf. Sci.*, vol. 282, pp. 111–135, 2014.

[23] J. Ahmad, F. Javed, and M. Hayat, "Intelligent computational model for classification of sub-Golgi protein using oversampling and fisher feature selection methods," *Artif. Intell. Med.*, vol. 78, pp. 14–22, 2017.

[24] H. Liu and L. Yu, "Toward integrating feature selection algorithms for classification and clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 17, no. 4, pp. 491–502, Apr. 2005.

[25] M. Dash *et al.*, "Feature selection for clustering-a filter solution," in *Proc. IEEE Int. Conf. Data Mining*, 2002, pp. 115–122.

[26] K. K. Bharti and P. K. Singh, "Hybrid dimension reduction by integrating feature selection with feature extraction method for text clustering," *Expert Syst. Appl.*, vol. 42, no. 6, pp. 3105–3114, 2015.

[27] M. Bennasar, Y. Hicks, and R. Setchi, "Feature selection using joint mutual information maximisation," *Expert Syst. Appl.*, vol. 42, no. 22, pp. 8520–8532, 2015.

[28] S. S. Shreem, S. Abdullah, and M. Z. A. Nazri, "Hybrid feature selection algorithm using symmetrical uncertainty and a harmony search algorithm," *Int. J. Syst. Sci.*, vol. 47, no. 6, pp. 1312–1329, 2016.

[29] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005.

[30] P. A. Estévez, M. Tesmer, C. A. Perez, and J. M. Zurada, "Normalized mutual information feature selection," *IEEE Trans. Neural Netw.*, vol. 20, no. 2, pp. 189–201, Feb. 2009.

[31] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," *in Proc. Int. Conf. Mach. Learn.*, 2001, vol. 1, pp. 577–584.

[32] L. V. D. Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, 2008.

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

[34] Q. Zhang and H. Li, "MOEA/D: A multiobjective evolutionary algorithm based on decomposition," *IEEE Trans. Evol. Comput.*, vol. 11, no. 6, pp. 712–731, Dec. 2007.

[35] 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 Bioinformat.*, vol. 9, no. 1, 2008, Art. no. 497.

[36] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[37] 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.

[38] C. A. C. Coello, G. T. Pulido, and M. S. Lechuga, "Handling multiple objectives with particle swarm optimization," *IEEE Trans. Evol. Comput.*, vol. 8, no. 3, pp. 256–279, Jun. 2004.

[39] B. Trawiski, M. Smętek, Z. Telec, and T. Lasota, "Nonparametric statistical analysis for multiple comparison of machine learning regression algorithms," *Int. J. Appl. Math. Comput. Sci.*, vol. 22, no. 4, pp. 867–881, 2012.

[40] S. Rostami, D. O'Reilly, A. Shenfield, and N. Bowring, "A novel preference articulation operator for the evolutionary multi-objective optimisation of classifiers in concealed weapons detection," *Inf. Sci.*, vol. 295, pp. 494–520, 2015.

[41] G. Deco, W. Finnoff, and H. G. Zimmermann, "Unsupervised mutual information criterion for elimination of overtraining in supervised multilayer networks," *Neural Comput.*, vol. 7, no. 1, pp. 86–107, 1995.

[42] K. Y. Yeung, D. R. Haynor, and W. L. Ruzzo, "Validating clustering for gene expression data," *Bioinformatics*, vol. 17, no. 4, pp. 309–318, 2001.

[43] L. J. Van't Veer *et al.*, "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[44] T. Barrett *et al.*, "NCBI GEO: Archive for high-throughput functional genomic data," *Nucleic Acids Res.*, D885–D890, 2009.

[45] W. L. Jamieson-Gladney *et al.*, "The chemokine receptor CX 3 CR1 is directly involved in the arrest of breast cancer cells to the skeleton," *Breast Cancer Res.*, vol. 13, no. 5, 2011, Art. no. R91.

[46] M. C. Louie *et al.*, "ACTR/AIB1 functions as an E2F1 coactivator to promote breast cancer cell proliferation and antiestrogen resistance," *Molecular Cellular Biol.*, vol. 24, no. 12, pp. 5157–5171, 2004.

[47] P. M. Campeau *et al.*, "Hereditary breast cancer: New genetic developments, new therapeutic avenues," *Human Genetics*, vol. 124, no. 1, pp. 31–42, 2008.

[48] X. Yang *et al.*, "CDKN1C (p57T KIP2) is a direct target of EZH2 and suppressed by multiple epigenetic mechanisms in breast cancer cells," *PLoS One*, vol. 4, no. 4, 2009, Art. no. e5011.

[49] L. Xu *et al.*, "GPR56 plays varying roles in endogenous cancer progression," *Clin. Exp. Metastasis*, vol. 27, no. 4, pp. 241–249, 2010.

[50] K. Milde-Langosch *et al.*, "Validity of the proliferation markers Ki67, TOP2A, and RacGAP1 in molecular subgroups of breast cancer," *Breast Cancer Res. Treatment*, vol. 137, no. 1, pp. 57–67, 2013.

[51] M. Ohta *et al.*, "Clinical significance of the reduced expression of G protein gamma 7 (GNG7) in oesophageal cancer," *Brit. J. Cancer*, vol. 98, no. 2, pp. 410–417, 2008.

[52] Y. Wang, T. Jin, X. Dai, and J. Xu, "Lentivirus-mediated knockdown of CEP55 suppresses cell proliferation of breast cancer cells," *Biosci. Trends*, vol. 10, no. 1, pp. 67–73, 2016.

[53] P. K. Khade and P. Giannakakou, "RABbing cancer the wrong way," *Proc. Nat. Acad. Sci.*, vol. 111, no. 31, pp. 11230–11231, 2015.