IEEE Journal of Translational Engineering in Health and Medicine

Received 26 September 2018; revised 3 December 2018 and 31 December 2018; accepted 3 January 2019. Date of publication 7 March 2019; date of current version 21 March 2019.

Digital Object Identifier 10.1109/JTEHM.2019.2891746

Nature-Inspired Multiobjective Cancer Subtype Diagnosis

YUNHE WANG[1], BO LIU[2], ZHIQIANG MA[1], KA-CHUN WONG[3], AND XIANGTAO LI[1]

[1]School of Information Science and Technology, Northeast Normal University, Changchun 130117, China
[2]School of Physical Education, Northeast Normal University, Changchun 130117, China
[3]Department of Computer Science, City University of Hong Kong, Hong Kong

CORRESPONDING AUTHOR: X. LI (lixt314@nenu.edu.cn)

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 20190103006JH, in part by the Science and Technology Development Planning of Jilin Province under Grant 20160204043GX, in part by the Fundamental Research Funds for the Central Universities under Grant 2412017FZ026, and in part by the Chongqing High-Performance Computing Platform: 991 cstc2015ptfw-ggfw120002.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author.

ABSTRACT Cancer gene expression data is of great importance in cancer subtype diagnosis and drug discovery. Many computational methods have been proposed to classify subtypes using those data. However, most of the previous computational methods suffer from poor interpretability, experimental noises, and low diagnostic quality. To address those problems, multiobjective ensemble cuckoo search based on decomposition (MOECSA) is proposed to optimize those four objectives simultaneously including the number of features, the accuracy, and two entropy-based measures: the relevance and the redundancy, classifying the cancer gene expression data with high predictive power for different cardinality levels under multiple objectives. A novel binary encoding is proposed to choose gene subsets from the cancer gene expression data for calculating four objective functions. Furthermore, an effective ensemble mechanism blended in the cuckoo search algorithm framework is applied to balance the convergence speed and population diversity in MOECSA. To demonstrate the effectiveness and efficiency of the proposed algorithm, experiments on thirty-five benchmark cancer gene expression datasets, four independent disease datasets, and one sequencing-based dataset are carried out to compare MOECSA with the six state-of-the-art multiobjective evolutionary algorithms and seven traditional classification algorithms. The experimental results in different perspectives demonstrate that MOECSA has better diagnosis performance than others at multiple levels.

INDEX TERMS Classification, feature selection, cancer subtype diagnosis, multiobjective optimization.

I. INTRODUCTION

Cancer diagnosis across gene expression data analysis has emerged as an active area of research over the past decades in medicine. High-throughput sequencing enables us to measure the related gene expression levels simultaneously [1]. Since the characteristics of the cancer gene expression data are high-dimensional, noisy, and sample-imbalanced, it is difficult to carry out the diagnosis task in an efficient way. Therefore, the demanding job on the cancer gene expression data is to develop effective methods for classifying the samples into subtypes accurately with a small subset of informative genes. It makes sense that feature selection [2] is considered as a necessary pretreatment process to analyze the cancer gene expression data for reducing the dimensionality of the data.

In fact, the main purpose of cancer diagnosis is to find the high predictive accuracy using small number of genes [3]. It means that these two specific objectives are maximizing the classification accuracy and minimizing the number of features in the subsets. Since the two objectives are conflicting, it would be better to treat the classification problem as a multiobjective problem rather than a single objective problem. In the past decades, a plethora of data classifying multiobjective approaches [4] have been proposed to trade off the discriminating power and the number of features. Ke *et al.* [5] presented a multi-objective ant colony optimization algorithm (MOACO) based on pareto dominance for selecting informative features and diagnosing the genes accurately and competitively. In [6], a multiobjective genetic algorithm (MOGA) was proposed, which aimed to search the

IEEE Journal of Translational Engineering in Health and Medicine

subsets of features effectively by combining different filter approach criteria. It made use of the general characteristics of the data to feature correlation. Bhattacharyya *et al.* [7] proposed a classification method based on archived multiobjective simulated annealing in order to predict miRNA promoters in the use of the classifier of SVM with RBF kernel. In [8], a multiobjective binary biogeography based optimization (MOBBO) and SVM with the leave-one-out cross-validation method used as a classifier were applied to optimize these two objectives simultaneously. In particular, MOBBO blended the non-dominated sorting method and the crowing distance with the BBO framework. However, these evolutionary algorithms always suffer from the unexpected balance between the exploration and exploitation, poor generalization ability, and too much computation time. Therefore, it is necessary to develop superior evolutionary algorithms to alleviate the shortcomings.

Cuckoo search algorithm (CSA) is a nature-inspired evolutionary algorithm imitating the behaviours of cuckoos, a kind of parasitic birds, for optimization problems developed by Yang and Deb [9]. Thanks to the fast convergence and the diversity in the distribution of solutions, it has been applied to many different real-world problems. What is more, numerous types of multiobjective cuckoo search algorithms are conducted on various research fields. Yang and Deb [10] proposed a multiobjective cuckoo search for design optimization in engineering. In [11], a cuckoo search algorithm was given to solve a multiobjective job shop scheduling problem using a pareto archive to keep all nondominated solutions. A multiobjective cuckoo search algorithm based on Duffing's Oscillator was introduced by Coelho *et al.* for Jiles-Atherton vector hysteresis parameters determination of hysteresis models [12]. Syberfeldt Anna proposed [13] a multiobjective cuckoo search to maximize machine utilizations and minimize the tied-up capital simultaneously in the real-world manufacturing process. Liang and Kwan [14] put forward a multiobjective cuckoo search algorithm to optimize the filter coefficients of FIR lowpass and bandpass digital filters. Moreover, a mutiobjective fractional cuckoo search was proposed by George *et al.* to cluster high dimensional data accurately [15]. Zhang *et al.* [16] proposed a hybrid multiobjective cuckoo search on benchmark MOPs of the multiobjective function optimization problem. Although, multiobjective cuckoo search algorithm has been applied on a variety of research fields so far, the multiobjective cuckoo search algorithm to solve the multiobjective classification problem is still in infancy. As a result, in this paper, the multiobjective ensemble cuckoo search algorithm based on decomposition (MOECSA) for cancer subtype diagnosis is proposed to optimize the four objective functions, including the number of feature, the accuracy and two entropy-based measures: the relevance and the redundancy simultaneously. Compared MOECSA with other existing multiobjective algorithms, the main new contributions of MOECSA can be summarized as follows:

- Based on the objective functions of each solution, a gene subset is required firstly. Hence a novel binary encoding method is proposed to select the gene subsets for calculating the fitness of potential solutions.
- Inspired by the differential evolutionary algorithm (DE), two improved search methods are proposed to trade off the exploitation and exploration based on the current individual and its neighbors.
- An effective ensemble mechanism is designed to help the algorithm extract the intrinsic complexity information from the cancer gene expression data. In this ensemble mechanism, those two improved search methods are updated with the help of the successful experience from the previous generations for searching highly qualified potential solutions.

In order to verify the performance of the multiobjective ensemble cuckoo search algorithm (MOECSA), experimental results are presented and compared with six state-of-the-art multiobjective evolutionary algorithms and seven classification algorithms. We also conduct the time complexity analysis, the parameter analysis, and extended experiments to demonstrate the efficiency and robustness of MOECSA from different perspectives.

II. METHODS

A. CUCKOO SEARCH ALGORITHM (CSA)

Cuckoo search [9] is a novel nature-inspired evolutionary algorithm by imitating the obligate parasitism behaviors of some cuckoos that lay their own eggs in the nest of other host birds for searching the optimal solutions. Firstly, each cuckoo can lay only one egg and drop its egg in a selected nest randomly. Therefore each egg can be treated as an individual (a solution). Then, the better egg with the better fitness value can enter to the next generation. After that, the egg of a cuckoo can be found by the host bird with a probability, which leads the laid egg to be thrown away or the host bird to build a new nest. The cuckoos search the entire decision space to find the optimal solutions by recording the fitness value of all the candidate solutions.

In the search process, there are a fixed number (NP) of nests (the initial population) placing in the search space. Besides, solutions in the population should cover the whole decision space as much as possible. The initial population $P = \{S_1, \ldots, S_{NP}\}, S_i = \left\{s_i^1, \ldots, s_i^D\right\} (i \in \{1, \ldots, NP\})$ is uniformly randomly chosen from the minimum and maximum bounds $S_{min} = \left\{s_{min}^1, \ldots, s_{min}^D\right\}$ and $S_{max} = \left\{s_{max}^1, \ldots, s_{max}^D\right\}$. Each individual within the search space at generation t is generated by:

$$s_i^k(t) = s_{min}^k + r \times (s_{max}^k - s_{min}^k) \quad (k \in \{1, 2, \ldots, D\}) \quad (1)$$

Where r is a random number generated in [0,1].

After that, in the cuckoo search the new solution $S_i(t+1)$ at $t+1$ generation for the cuckoo $S_i(t)$ $(i \in \{1, \ldots, NP\})$ is generated by the Lévy flight in the search process.

A Lévy flight is employed as follows:

$$S_i(t+1) = S_i(t) + \alpha \oplus Lévy(\beta) \tag{2}$$

Where the product $\oplus$ represents entry-wise multiplications. The Lévy flight shown as below is a random walk that observes a Lévy distribution, which has an infinite variance with an infinite mean. Thus the consecutive steps of a cuckoo generate a random walk process that obeys a power law length distribution with a heavy tail. The process of producing new solutions can be also regarded as a Markov chain by a stochastic equation for random walk.

$$Lévy \sim \mu = t^{-1-\beta} \quad (0 < \beta \leqslant 2) \tag{3}$$

Besides, $\alpha > 0$ is a real number indicating the step size and it should be related to the scales of the problem of interest. Since the step size using the Lévy flight is not trivial, a simple scheme [9] for producing S_{new_i} can be calculated as following equations:

$$\alpha = \alpha_0 \times Step_i \times (S_i - S_j) \quad (i, j \in \{1, \ldots, NP\}) \tag{4}$$

$$Step_i = \frac{\mu_i}{\left|v_i^{1/\beta}\right|} \quad (i \in \{1, \ldots, NP\}) \tag{5}$$

$$\mu \sim N(0, \sigma_\mu^2), \quad v \sim N(0, \sigma_v^2) \tag{6}$$

$$\sigma_\mu = \left\{\frac{\tau(1+\beta)sin(\pi\beta/2)}{\tau[(1+\beta)/2]\beta 2^{(\beta-1)/2}}\right\}^{1/\beta}, \quad \sigma_v = 1 \tag{7}$$

$$S_{new_i} = S_i + \alpha \times randn\,[D] \tag{8}$$

Where τ is the standard Gamma function and randn[D] is a standard normal distribution with the size [1,D]. α_0 can range from 0.01 to 0.5. The Lévy exponent β can be 0.5, 1, 1.5, and 2.

After generating the new individual S_{new_i}, the objective value of S_{new_i} is compared with the fitness of S_i. If the fitness of S_{new_i} is better than that of S_i, S_{new_i} replaces S_i and is accepted as a new individual to enter to the $t+1$ generation. Otherwise, S_i is retained in the population.

Then, a fraction ($p_a \in (0, 1)$) of nests can be found by host birds. As a result, a simple way to build a new nest S_{new_j} replacing the discovered nest S_j can be described as:

$$S_{new_j} = S_j + r \times (S_{j_1} - S_{j_2}) \tag{9}$$

Where r is a random number drawn from [0,1] and j_1, j_2 are random integers in the range [1,NP]. The crossover operator to select nests is performed as follows, where $rand_j$ is a random number ranged from 0 to 1:

$$S_j(t+1) = \begin{cases} S_{new_j} & If\ rand_j < pa \\ S_j(t) & If\ rand_j \geqslant pa \end{cases} \tag{10}$$

When the egg of a cuckoo is much more similar to the egg of a host bird, it is more difficult to be found in the real world. Hence it is worth noting that the main reason to use a random walk with some step sizes randomly is that the objective function value is closely connected with the difference between the host bird's egg and the cuckoo's egg.

In this paper, to understand the variables in the cuckoo search algorithm easily, a nest or an egg represents a solution or an individual. A solution indicates a gene representation chosen from the genes of the cancer gene expression data using the binary encoding. The genes of the cancer gene expression data are denoted as the search area. Therefore, a population includes NP individuals is generated to conduct the proposed algorithm.

B. MULTIOBJECTIVE CLASSIFICATION BY ENSEMBLE CUCKOO SEARCH ALGORITHM (MOECSA)

1) MULTIOBJECTIVE OPTIMIZATION

In multiobjective optimization problems, there are multiple (two or more) various and conflicting objectives being optimized together. In mathematics, a multiobjective optimization problem can be defined as follows:

$$min\,\{f_1(x), \ldots, f_M(x)\} \quad (x \in X) \tag{11}$$

Where M is the number of objectives, x is a D-dimensional decision vector with D variables, X is the decision space of all the available decision vectors. Besides, multiobjective optimization is concerned with some important concepts that are *dominance*, *pareto set (PS)*, and *pareto frontier (PF)*. For instance, in a minimization problem, if a feasible solution x_1 can dominate another feasible solution x_2, then only if $f_i(x_1) \leqslant f_i(x_2)$ for each i, and there exits at least one i, $f_i(x_1) < f_i(x_2)$ ($i \in \{1, \ldots, M\}$). A set of all the nondominated solutions is called a pareto set. Then, a nondominated point is an image of the objective vector in terms of a nondominated solution in objective space. The pareto frontier, which exhibits different tradeoff curves of the conflicting objectives, is a set of all nondominated points.

2) OBJECTIVE FUNCTIONS

A suitable choice of objective functions takes important part in multiobjective classification. The tradeoff between the number of features (cardinality) and the classification performance of the model formed by the selected features has been an emerging trend for multiobjective classification. However, the two entropy-based measures namely the relevance and the redundancy for selected features are often ignored [17]. The measure of relevance is employed to assess the discriminating power of the chosen features and the redundancy measures the level of similarity among them [18].

Under a comprehensive consideration of these four objective functions, the cancer gene expression data can be more adequately interpreted than that using the two objectives. This algorithm with four objectives makes sense to find a set of subsets having high predict power for different cardinality levels. In this paper, let x_i is a set of candidate features/genes, y is the target label, a subset of x_i is call X, those objectives

IEEE Journal of Translational Engineering in Health and Medicine

can be defined as follows:

$$f_1(X) = max\left\{\sum_{x_i \in X} SU(x_i, y)\right\} \tag{12}$$

$$f_2(X) = min\left\{\sum_{x_i, x_j \in X, i<j} SU(x_i, x_j)\right\} \tag{13}$$

$$f_3(X) = min\{|X|\} \tag{14}$$

$$f_4(X) = max\left\{\frac{tp + tn}{tp + tn + fp + fn}\right\} \tag{15}$$

Where $SU(a, b)$ stands for the symmetric uncertainty between a and b [17], [19]. tp, tn are true positives, true negatives. On contrary, fp, fn represent false positives, false negatives.

3) MULTIOBJECTIVE ENSEMBLE CUCKOO SEARCH ALGORITHM FOR CANCER SUBTYPE DIAGNOSIS

In this section, we raise a multiobjective algorithm to blend the improved CSA framework with an ensemble mechanism regarding to the four objectives for diagnosing the cancer data. The framework of MOECSA and the effective subdivision techniques are stated below in detail.

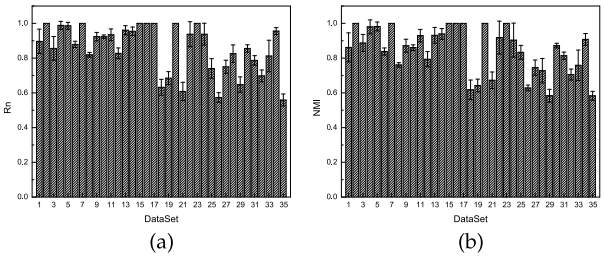
a: Structure of MOECSA

In MOECSA, the tchebycheff approach is employed to decompose this multiobjective classification problem with four objectives into a number of scalar classification subproblems. It is a less expensive approach computationally to solve multiobjective problems. In detail, we apply a weight vector $\lambda = \{\lambda^1, \ldots, \lambda^M\}$, and $\sum_{k=1}^{M} \lambda^k = 1$, $\lambda^k \geqslant 0$ (M is the number of objectives) on the proposed algorithm to compute the single objective function of a subproblem shown in Eqs.(16).

$$g^{te}(x|\lambda, z^*) = \max_{1 \leqslant k \leqslant M}\left\{\lambda^k \left|f_k(x) - z_k^*\right|\right\} \tag{16}$$

where $z^* = \{z_1^*, \ldots, z_M^*\}$ is the reference point. z_k^* is the best value of each objective $f_k(x)$ found at present [20]. Besides, if there exists NP weight vectors $\{\lambda_1, \ldots, \lambda_{NP}\}$ and each weight vector has M dimensions, a cancer subtype diagnosis problem with M objectives is divided into NP cancer subtype diagnosis subproblems. Each weight vector $\lambda_i^j (i \in \{1, 2, \ldots, NP\}, j \in \{1, 2, .., M\})$ is assigned to a subproblem with M objectives. By Eqs.(16), each subproblem i is integrated into a single objective $g^{te}(x|\lambda_i^j)$ to update the individuals in the population. The main loop of MOECSA is provided by Supplementary Algorithm S1. It reflects a multiobjective evolutionary algorithm consisting of *Get a new nest section* and *The empty nests section* in Supplementary Algorithm S2 and S3 respectively.

As to the initial section, NP individuals, each of which is a D-dimensional vector with random numbers from $[-4, 4]$, are initialized to generate a population $P = \{S_1, \ldots, S_{NP}\}$ in MOECSA. It is noted that we propose to use the binary coding



(a) (b)

FIGURE 1. Performance of MOECSA on 35 benchmark datasets for R_n in (a) and *NMI* in (b). The horizontal axis denotes different datasets while the vertical axis denotes the mean with standard deviation.

to transfer NP individuals into binary strings for evaluating the objective functions, but the individuals encoded in decimal numbers are adopted to evolve the population. NP subproblems are generated corresponding to NP subpopulations and each subpopulation has four objectives. Aggregating the four objectives using Eqs.(16), the single objective fitness is evaluated for each subproblem. In the search process, the solution is kept or rejected for the next generation based on its single objective value. Next, different weight vectors are assigned to each subproblem and $E_i = \{i_1, \ldots, i_T\} (i \in \{1, \ldots, NP\})$ is defined in terms of the Euclidean distance between its vector and other weight vectors. The T closest weight vectors are created to explore the better subpopulations in the neighborhood region and update the fitness of the subproblem iteratively. In each iteration, at first for each individual, *Get a new nest section* (Supplementary Algorithm S2) is employed to build a new solution by Lévy flight. It is necessary to point out that two individuals are chosen randomly from T neighbourhoods. After that, the single objective composed of the four objectives of the new solution using λ weight vector is compared with all the neighborhood individuals. If the new solution S_{new_i} is superior to the neighborhood $S_{E_i^j}$, then $S_{E_i^j}$ and its fitness are replaced by S_{new_i} and better fitness value respectively. An improved population including NP individuals with better fitness is generated by iterations. After that, the ensemble mechanism is provided to assign S_i with a designated p_a value and a search strategy. At the last section, we apply *The empty nests section* (Supplementary Algorithm S3) utilizing the improved search methods to abandon old nests and build new ones with p_a value. If S_{new_i} performs better than S_i in the fitness of the single objective, S_i can be replaced by S_{new_i}, thus the utilized p_a value and the search strategy are added to the candidate pool. In addition, the number of count increases by one and we adopt the count number to calculate the selection rate for selecting the corresponding p_a and the search method in the next generation. All in all, a population with NP high quality individuals replaced by better fitness individuals is produced. In these two sections, for each objective the reference point is updated once the fitness of S_{new_i} is less than z_i^*. At last, a pareto set including all nondominated individuals is given.

b: Binary encoding

In this paper, considering that a binary individual is required when calculating the fitness of objective functions, a new binary encoding is proposed to transfer an individual with the continuous encoding to the binary encoding. To carry out the binary encoding, we adopt the mutual information algorithm [21] to select D (200) important features first. In MOECSA, each individual is an D-bit binary string where D is the number of features in the subset. For the bit value, "1" represents that the feature is chosen from the subset and otherwise "0". The binary encoding strategy is expressed as follows:

$$h = \left| \frac{2}{\pi} arctan(\frac{\pi}{2} s_i^k(t)) \right| \tag{17}$$

$$BinaryCode(s_i^k(t)) = \begin{cases} 0 & (If\ \varphi < h) \\ 1 & (If\ \varphi \geqslant h) \end{cases} \tag{18}$$

Where φ is drawn from a Gaussian distribution which is $\varphi \sim N(0.5, 0.1^2)$, $s_i^k(t)$ is a decimal dimension of an individual vector of the population.

c: Boundary constraints

During the search of MOECSA, if some individuals denoted as decimal vectors move out of the search space bounds and become infeasible, the individual is assigned to a new value within the isolated and finite space using the following reset rule. It benefits from the repaired value to keep the population diversity instead of trapping in the local optimum using the boundary value replacement rule to some extent.

$$s_i^k = \begin{cases} min\left\{s_{max}^k, 2s_{min}^k - s_i^k\right\} & If\ s_i^k < s_{min}^k \\ max\left\{s_{min}^k, 2s_{max}^k - s_i^k\right\} & If\ s_i^k > s_{max}^k \end{cases} \tag{19}$$

Where $i \in \{1, \ldots, NP\}$ and $k \in \{1, \ldots, D\}$.

d: Improved search methods

In the standard cuckoo search algorithm, the second part of algorithm abandons cuckoos by Eqs.(9) and Eqs.(10). Inspired by DE [22], two novel search mechanisms and a simple crossover operator are proposed to increase the performance of MOECSA. DE is an effective population_based stochastic search method with a simple structure. It exhibits remarkable performance in a variety of problems. Three operators: mutation, crossover, and selections form the basic framework of DE. Many different strategies are used in different domains due to the ability to enhance the diversity and speed up the convergence by generating a new trial vector. Therefore, based on DE and the property of CSA [23], two new modified search methods and a crossover operator according to the fraction p_a are employed when some nests of cuckoos are necessary to be emptied.

On one hand, to enhance the population diversity on breadth, a new search strategy expressed as Eqs.(20) is used. On the other hand, the technique shown as Eqs.(21) is employed to improve the exploitation capability of searching new optimal solutions. The illustration of the two search methods to produce a new trial vector respectively in minimization optimization with two objectives is summarized in Fig. 2.

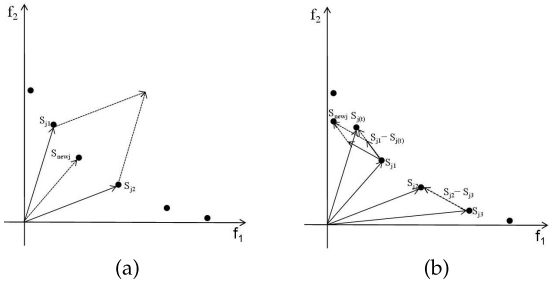


FIGURE 2. Illustration of two search methods. (a) Illustration of Eqs.(20). (b) Illustration of Eqs.(21).

$$S_{new_j}(t) = r_1 \times (S_{j_1} - S_j(t)) + r_2 \times (S_{j_2} - S_{j_3}) \tag{20}$$

$$S_{new_j}(t) = (1 - r_1) \times S_{j_1} + r_1 \times S_{j_2} \tag{21}$$

Where r_1 is a random number chosen from [0,1] and r_2 is a number drawn from a Gaussian distribution with the standard deviation "0.01" and the mean "0.1". $S_{j_1}, S_{j_2}, S_{j_3}$ are mutually exclusive individuals selected randomly from the current population and $S_j(t)$ is the current individual.

In addition, the crossover operator combined the search strategy with the abandon fraction p_a, for the current individual generating a target individual embedded into the empty nests section of MOECSA is defined below. It can get a balance on the exploration and exploitation of searching good quality solutions. Meanwhile this operator can make a good preparation for the follow ensemble mechanism.

$$S_{new_j}^k(t+1) = \begin{cases} S_{new_j}^k(t) & If\ rand_j^k < p_{aj} \\ S_j^k(t) & If\ rand_j^k \geqslant p_{aj} \end{cases} \tag{22}$$

e: The ensemble mechanism

In this part, the ensemble mechanism is proposed to select the suitable search method and its corresponding p_a value in the proposed algorithm. Since the candidate search methods in the pool should have distinctive characteristics to exhibit distinct performance capability in the entire search process. Therefore we choose the above two effective search methods in the pool since their distinct performance characteristics. Then, the value pool of p_a is taken in the range from 0.4 to 0.7 in the step of 0.1 to get appropriate solutions. If the p_a is too small then the search strategy poses little effect on searching unexplored solution regions, while if the p_a is a little large then a restart operator may be performed randomly.

As depicted in Supplementary Algorithm S1, for the ensemble mechanism, we assign each individual to a p_a value and a search method randomly chosen from distinct pools. At first iteration, the p_a value and the search method are generated from the respective pools randomly. At other iterations, if a random number generated between 0 and 1 is less than the selective probability and the candidate pool is not empty, we adopt the p_a value and the search strategy chosen randomly from the candidate pool. Otherwise,

the same method to generate those two components is applied as the first iteration. After that, the selective probability is computed by the average value among the better individuals count percentage of *NP* individuals in each iteration for all the current iterations. Besides, the better p_a value and search strategy are added to a candidate pool based on the selective probability. In conclusion, we assign the p_a value and the search strategy to each individual chosen either from the combination candidate pool according to the replacement probability or the respective pools randomly. In this way, it is noted that the better p_a and search method can pass to the next generation for building high quality solutions with an increased probability. Meanwhile, it is also responsible for diversity because of the selection randomness in the corresponding pools. In addition, time complexity analysis is provided in the first section of Supplementary.

C. EXPERIMENTS

1) DATA SOURCES

In this paper, 35 cancer gene expression datasets are employed as the benchmark datasets. The truth label information of these datasets are all achieved from [24]. Supplementary Table S1 shows these 35 benchmark cancer gene expression datasets [25], [26]. As demonstrated in Supplementary Table S1, the number of samples is ranged from 22 to 248 for all benchmark datasets; the number of features/genes is varied from 85 to 4553 in the data and the number of classes is varied from 2 to 14. Besides, from Supplementary Table S1, we see that some datasets are from the same data source. For example, Alizadeh-2000-v2 and Alizadeh-2000-v3 use the same source which the last one has one more class than the first one; Golub-1999-v1, Golub-1999-v2 and Yeoh-2002-v1, Yeoh-2002-v2 are also from the same source with different number of classes respectively; Armstrong-2002-v1 and Armstrong-2002-v2 have the same number of samples; Tomlins-2006-v1 and Tomlins-2006-v2 are the same as Lapointe-2004-v1 and Lapointe-2004-v2 which they have distinct numbers of samples and genes but the first one has one more class respectively.

2) PARAMETER SETTING

35 cancer gene expression datasets are used to compare the performance of different algorithms. In terms of MOECSA, when producing a set of weight vectors, each component of a weight vector selects the values from $\{(0/H), \ldots, (H/H)\}$, where H is an integer setting to 7 and the number of weight vectors $NP = C_{M-1}^{H+M-1}$ is 120, where $M = 4$ is the number of objectives. And the neighborhood weight vector number T for each weight vector is 50. The essential components α_0 and β for Lévy flight are set to 0.1 and 1 respectively. They are all the best algorithm settings discussed in Section 2.3.7. To show a reasonable comparison, the number of fitness evaluation is set as the stopping criterion rather than generation times or CPU time. We set 1000 times of objective function evaluation (FE) for each dataset to run each algorithm. At the same time, for achieving the statistical significance, each multiobjective evolutionary algorithm runs 30 times independently on each dataset. Therefore, an average result on 30 independent runs is calculated to analyze the performance of MOECSA on each cancer gene expression dataset.

3) EVALUATION METRICS

In order to measure the diagnosis results, we compare the labels obtained by the classification algorithms with the truth labels of all the samples. The normalized mutual information (*NMI*) and normalized rand index (R_n) are used to evaluate the diagnosis performance of all the compared algorithms, which can compute the similarity between the classification labels and the ground truth labels. Therefore the results of these two metrics with higher values are better than other results in cancer subtype diagnosis.

NMI is ranged from 0 to 1, which represents no mutual information (MI) to a perfect correlation, for presenting the normalization of MI. It can be computed as follows:

$$NMI = \frac{\sum_{i,j} n_{i,j} log \frac{n \cdot n_{i,j}}{n_{i+} \cdot n_{+j}}}{\sqrt{(\sum_i n_{i+} log \frac{n_{i+}}{n})(\sum_j n_{j+} log \frac{n_{+j}}{n})}} \tag{23}$$

R_n is expressed as the selection index to evaluate the agreement between two categories, which can be calculated as follows:

$$R_n = \frac{\sum_{i,j} \binom{n_{i,j}}{2} - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{\sum_i \binom{n_{i+}}{2}/2 + \sum_j \binom{n_{+j}}{2}/2 - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}} \tag{24}$$

4) OTHER RELATED METHODS FROM LITERATURE

In one aspect, to validate the performance of multiobjective algorithms, multiple effective multiobjective evolutionary algorithms are compared with our proposed algorithm. They are nondominated sorting genetic algorithm II (NSGA-II) [4], multiobjective differential evolution (MODE) [27], region based pareto envelope based selection algorithm (PESA-II) [28], multiobjective particle swarm optimization (MOPSO) [29], grid-based evolutionary algorithm (GrEA) [30], and hypervolume-based algorithm (HypE) [31]. In order to demonstrate the effectiveness of MOECSA, they represent different algorithmic paradigms. Under the perspective of multiobjective evolutionary optimization algorithms, the proposed algorithm MOECSA is a multiobjective algorithm built on the foundation of decomposing a multiobjective problem into several single objective optimization problems. NSGA-II adopts strategies including the nondominated sorting method and the crowding distance strategy. MODE applies the differential evolution algorithm to the multiobjective problems. PESA-II is a multiobjective evolutionary algorithm using the mechanism of genetic algorithm with region-based selection based on pareto envelope. Besides, MOPSO is a multiobjective algorithm using the strategy of sharing information and moving towards global best particles and their own local best memory like PSO.

GrEA and HypE are grid-based and hypervolume-based algorithms respectively.

In another aspect, several traditional classification methods including K-nearest neighbors algorithm (KNN) [32], extreme learning machine (ELM) [33], support vector machine (SVM) [34], Baysian classification, generalized learning vector quantization classification (GRLVQ), Adaboosting, and artificial neural network (ANN) [35] are employed to compare the efficiency with our proposed algorithm. Under the perspective of different classification algorithms, they indicate different classification theories. They are all supervised learning algorithms. KNN is a simple classification algorithm on the basis of measuring the distance between different eigenvalues which is adopted in MOECSA to obtain the classification accuracy. ELM is a learning algorithm based on the random input weight and hidden layer biases for classifying. SVM uses a model that is finding the best separating hyperplane in the feature space to maximize the intervals between the positive and negative samples on the training set. Baysian is a classification algorithm based on the Bias theorem. GRLVQ applies a winner-take-all Hebbian learning-based approach on classification. Adaboosting is a kind of boost method using a series of classifiers. ANN uses a nonlinear adaptive information processing system including a large number of processing unit to classify samples.

For gaining empirical insight into the comparison in statistic, we compute the paired Wilcoxon's signed rank test to measure the performance between pairs of algorithms significantly. There are three symbols "$-$", "$+$", and "$\approx$" to express this typical nonparametric statistical hypothesis test method. The "$+$" indicates that our proposed algorithm performs better than the other algorithms while the "$-$" is shown the compared algorithm is superior to our algorithm. The "$\approx$" denotes that the compared algorithm is not different significantly in terms of the other algorithm. The p-Value with less than 0.05 means that there is a significant comparison between the two algorithms and then it is worthy to carry out this test.

5) EVALUATION ON BENCHMARK CANCER GENE EXPRESSION DATA

This section is designed to investigate the better performance of MOECSA from the multiobjective algorithm perspective by comparing with different multiobjective evolutionary algorithms. Six existing state-of-the-art algorithms on 35 benchmark cancer gene expression datasets are employed including NSGA-II [36], MODE, PESA-II, MOPSO, GrEA, and HypE in order to validate the effectiveness of MOECSA.

To compare different multiobjective evolutionary algorithms fairly, the classifier KNN with 10-fold cross-validation to evaluate the accuracy and the binary encoding proposed in this paper are adopted in the same way in these algorithms. Each algorithm runs 30 times independently on each cancer gene expression dataset. Supplementary Table S2 and S3 conclude the comparative results of those seven algorithms. In the last row of each table, it is given the statistical results by Wilcoxon's signed rank test. It is worthy noting that Fig. 1 shows the robustness of MOECSA on 35 benchmark datasets. From Fig. 1, in terms of both R_n and *NMI*, it is concluded that MOECSA exhibits high robustness in cancer subtype diagnosis on 35 cancer gene expression datasets.

1) For the evaluation metrics R_n, it can easily be found that MOECSA presents the better performance than other algorithms as shown in Supplementary Table S2. Several observations can be concluded as follows. a) The proposed algorithm MOECSA outperforms other algorithms in most datasets while NSGA-II provides the worst solutions. b) MOECSA can obtain the best R_n result "1" on seven datasets numbered 2, 7, 15, 16, 17, 20, 23, with the same performance of HypE in getting the best R_n result "1". For other algorithms NSGA-II, MODE, PESA-II, MOPSO, and GrEA, there are 1, 2, 3, 1, and 5 best R_n results respectively. c) MOECSA is worse than HypE on Garber-2001, Risinger-2003 and GrEA on Bredel-2005 and Garber-2001. For Armstrong-2002-v1, HypE obtains the same result with MOECSA. d) MOECSA provides the better results on 34, 33, 32, 34, 28, and 25 datasets compared with NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE respectively according to the statistical results on the last row of Supplementary Table S2. e) Fig. 3(a) shows the diagnosis performance of different multiobjective evolutionary algorithms on 35 cancer gene expression datasets which is measured by R_n boxplot. It illustrates a better overall performance of MOECSA compared with other multiobjecitve algorithms intuitively.

2) For the evaluation metrics *NMI*, we can find that Supplementary Table S3 shows different performance of different multiobjective evolutionary algorithms. Our proposed algorithm MOECSA surpasses other algorithms obviously. From the average results running 30 times on each dataset in Supplementary Table S3, several observations can be provided. a) The proposed algorithm MOECSA can obtain the best results among all the algorithms in most datasets while NSGA-II obtains the worst results. b) MOECSA can give the best *NMI* result "1" on seven datasets, namely Alizadeh-2000-v2, Bittner-2000, Gordon-2002, Khan-2001, Laiho-2007, Liang-2005, and Nutt-2003-v3. The performance of it is equal to HypE in getting best *NMI* result "1". Regarding to other algorithms, NSGA-II and MOPSO can both obtain the best *NMI* result "1" on Nutt-2003-v3; MODE, PESA-II, GrEA can obtain the best *NMI* result "1" on 2, 3, 5 datasets respectively. c) MOECSA is worse than GrEA on Bredel-2005 and Garber-2001. For HypE, it is better than MOECSA on Garber-2001. In terms of Armstrong-2002-v1, HypE and MOECSA can get the same result 0.9801. d) MOECSA can give better results on 34, 33, 32, 34, 28, and 26 datasets compared with the other algorithms, NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE respectively on the basis of the last row of Supplementary Table S3. e) Fig. 3(b) demonstrates the cancer subtype diagnosis performance of different multiobjective evolutionary algorithms measured by *NMI* boxplot. It is clear that better NMI values are distributed centrally across the 35 cancer gene expression datasets in MOECSA.

IEEE Journal of Translational Engineering in Health and Medicine

To show the level of the algorithm in minimizing the number of genes in cancer subtype classification, the compared results about the accuracy and the number of genes of different multiobjective evolutionary algorithms, including NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE, are shown in Supplementary Table S13 and Supplementary Table S14 respectively. We apply 35 cancer gene expression datasets to each algorithm and each algorithm runs 30 times independently on each dataset. Meanwhile, the statistical results by Wilcoxon's signed rank test are concluded in the last row of each table. For the accuracy, from Supplementary Table S13, MOECSA performs better accuracies than NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE on 31, 33, 32, 34, 30, 28 datasets out of 35 cancer gene expression datasets respectively. In terms of NSGA-II, it outperforms MOECSA in Lapointe-2004-v2, Ramaswamy-2001, Su-2001, and Yeoh-2002-v2. For other algorithms, they perform equal to MOECSA on 2, 3, 1, 5, and 7 datasets respectively. In addition, Supplementary Fig. S5 shows the overall performance of multiobjective algorithms intuitively by boxplots. For the number of genes, from Supplementary Table S14, regarding to NSGA-II, GrEA, and HypE, there is an apparent enhancement in minimizing the number of genes between the algorithm and the proposed algorithm. However, compared with MODE, PESA-II, and MOPSO, our proposed algorithm is slightly better than them in minimizing the number of genes. Since four objectives are adopted in our proposed algorithm, it is still a big improvement in decreasing the number of genes. Moreover, from Supplementary Fig. S6, an overall performance of cardinality using boxplots is given to demonstrate the level of minimization the number of genes for multiobjective algorithms. It can be seen that MOECSA can achieve less number of genes in cancer diagnosis on the 35 cancer gene expression datasets compared with other effective multiobjective algorithms.

Compared with GrEA and HypE, there are two main advantages of our proposed algorithm MOECSA. On one hand, in terms of minimizing the number of genes, MOECSA is superior to GrEA and HypE obviously according to Supplementary Table S14 and Supplementary Fig. S6. It is apparent that MOECSA has high ability to decrease the number of genes in cancer subtype diagnosis compared with GrEA and HypE. On the other hand, MOECSA has much simpler framework and more time-saving than GrEA and HypE according to the time complexity analysis provided in the first section of Supplementary. As GrEA is a grid-based algorithm and HypE is a hypervolume-based algorithm, much computation time has been consumed when calculating the grid-dominance relation and hypervolume values respectively. However, MOECSA is inspired by the simple structure of cuckoo search algorithm and it is based on decomposition, thus it saves much time in computing the dominant relationship. In summary, MOECSA outperforms GrEA and HypE not only in the performance of four objectives but also in time computation.

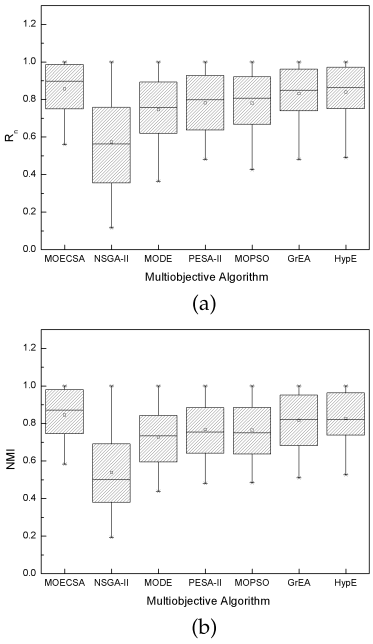


FIGURE 3. Comparison performance of different multiobjective evolutionary algorithms on 35 benchmark datasets. The performance is measured by R_n in (a) and *NMI* in (b).

Based on the experimental results, it is claimed that the proposed algorithm MOECSA can produce better diagnosis performance with high efficiency than other multiobjective evolutionary algorithms on 35 cancer gene expression datasets.

6) COMPARED WITH OTHER CLASSIFICATION ALGORITHMS

In this section, MOECSA is employed to classify cancer gene expression data into different subtypes. We compare our proposed algorithm with seven classification algorithms, namely KNN, ELM, SVM, Baysian, GRLVQ, Adaboosting, and ANN. 35 cancer gene expression datasets are adopted to test the performance of MOECSA. We run each algorithm 30 times independently on each dataset. The experimental results are provided in Supplementary Table S4 and S5. Besides, the last rows of Supplementary Table S4 and S5 summarize the statistical results by the Wilcoxons signed rank test.

1) For the evaluation metrics R_n, Supplementary Table S4 shows the R_n results of eight effective algorithms. Each method carries out 30 runs independently. As provided in Supplementary Table S4, multiple observations can be summarized. a) MOECSA is superior to KNN, ELM, SVM, Baysian, GRLVQ, Adaboosting, and ANN on 34, 34, 34, 35, 35, 32, and 35 datasets respectively. b) Adaboosting outperforms our proposed algorithm on Ramaswamy-2001, Singh-2002, and Yeoh-2002-v2. For Yeoh-2002-v1, SVM performs the best among all the compared classification algorithms. c) MOECSA can get the best result "1" on seven datasets numbered 2, 7, 15, 16, 17, 20, and 23. For KNN, it can only get one best result "1" on Khan-2001. While ELM obtains

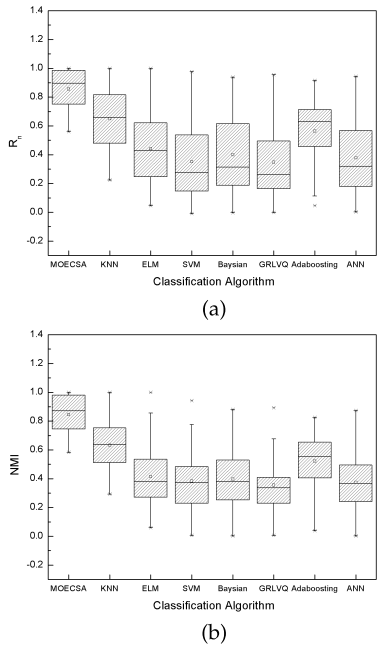


FIGURE 4. Comparison performance of different classification algorithms on 35 benchmark datasets. The performance is measured by R_n in (a) and *NMI* in (b).

one best result "1" on Nutt-2003-v3. d) Fig. 4(a) plots the diagnosis performance of different classification algorithms with respect to R_n on 35 cancer gene expression datasets using boxplots. A better overall performance in R_n of MOECSA compared with other traditional classification algorithms is provided by Fig. 4(a) intuitively.

2) For the evaluation metrics *NMI*, Supplementary Table S5 proves the high effectiveness of MOECSA on 35 cancer gene expression datasets with the average *NMI* results carrying out 30 independent runs for each classification algorithm. a) MOECSA can produce the best results on most datasets of all the compared classification algorithms. It is superior to other classification algorithms KNN, ELM, SVM, Baysian, GRLVQ, Adaboosting, and ANN, on 34, 34, 34, 35, 35, 34, and 35 datasets respectively. b) In terms of KNN and ELM, they can obtain the best result "1" on Khan-2001 and Nutt-2003-v3 respectively. While our proposed algorithm produces the best result "1" on Alizadeh-2000-v2, Bittner-2000, Gordon-2002, Khan-2001, Laiho-2007, Liang-2005, and Nutt-2003-v3. It is demonstrated that MOECSA has better performance in generating the best results "1". c) Concerning the inferior performance of MOECSA, SVM outperforms MOECSA on Yeoh-2002-v1 and Adaboosting performs better than MOECSA on Yeoh-2002-v2 respectively. d) Fig. 4(b) shows the diagnosis performance of different classification algorithms for *NMI* across 35 benchmark datasets with a good visualization by boxplots. From Fig. 4(b), it can be directly found that MOECSA beats other compared classification algorithms with better overall *NMI* results.

As evidenced by the above analyses, MOECSA is a highly competitive and effective multiobjective algorithm for cancer subtype diagnosis on thirty-five benchmark cancer gene expression datasets among the multiple traditional classification algorithms.

7) PARAMETER ANALYSIS

a: Effect of T in MOECSA

T is an indispensable parameter for MOECSA. To validate the sensitivity to T in the proposed algorithm on 35 cancer gene expression datasets, a T value set {30, 50, 70, 90} of different settings is provided for each dataset. An average *NMI* with 30 runs on each dataset is used to test the performance of different T values. Results of the average *NMI* are summarized in Supplementary Table S6 and Supplementary Fig. S1. As exhibited clearly in Supplementary Fig. S1, T=50 can produce high quality results statistically on these datasets in average.

b: Effect of Lévy exponent β in MOECSA

This section provides the comparative experiments by different settings of β in Lévy flight. Thanks to the description of the range for β in Section 2.1, β is varied over the set {0.5, 1, 1.5, 2} usually. We measure the results for each setting with the average *NMI* generating by 30 MOECSA runs on 35 cancer gene expression datasets. The results are concluded in Supplementary Table S7 and Supplementary Fig. S2 regarding to the average *NMI*. As shown in Supplementary Fig. S2, β=1 works well for MOECSA on these datasets with a better average *NMI* result.

c: Sensitive of α_0 in MOECSA

α_0 is an essential parameter to Lévy flight in generating new individuals randomly. Because of the range limitation about α_0 in Section 2.1 in usual, we select α_0 from a decimal set {0.05, 0.1, 0.15, 0.2}. Using the average *NMI* evaluated on these 35 datasets with 30 independent runs, Supplementary Table S8 and Supplementary Fig. S3 summarize the comparative performance by experiments. It can be easily observed from Supplementary Fig. S3 that α_0=0.1 is superior to the other settings for the 35 cancer gene expression datasets with the higher average *NMI*.

d: Sensitive of D in MOECSA

To investigate how number of genes used in the model can effect the performance of the proposed algorithm, this section gives the comparative experiments using different numbers of D, which is the pre-selected number of dimension to carry out the binary encoding. The number of the feature ranges from the set {50, 100, 150, 200, 300, 400, 500}. We use the average *NMI* on 35 cancer gene expression datasets to evaluate the performance of each number in the set. Each dataset runs 30 times independently. The results are summarized in Supplementary Table S12 and Supplementary Fig. S4. From Supplementary Table S12, it is shown that $D = 200$ can obtain a better overall performance with better average *NMI* across all the datasets for MOECSA. Meanwhile, we can find that various D maybe produce different performance for each dataset; for instance, for $D = 50$, it can

IEEE Journal of Translational Engineering in Health and Medicine

obtain better result on Alizadeh-2000-v1, Singh-2002, Yeoh-2002-v1, Yeoh-2002-v2. $D = 300$ can produce better results on Bhattacharjee-2001, Bredel-2005, Nutt-2003-v1, Tomlins-2006-v1. While for $D = 400$ and $D = 500$, they have better performance on Armstrong-2002-v2, Chowdary-2006, Lapointe-2004-v2, Pomeroy-2002-v1 and Armstrong-2002-v1 respectively. Therefore, we can conclude that $D = 200$ is not always the best setting for all datasets. From Supplementary Fig. S4, it is also shown that a *NMI* developing trend is generated across the 35 benchmark datasets. With the decrease of the number D, some informative features could be eliminated and the smaller features cannot cover all the dataset. Conversely, when D becomes larger, the algorithm could not effectively choose the informative features from the numerous features. Therefore, $D = 200$ is more suitable than other settings for the proposed algorithm.

D. EXTENDED PERFORMANCE COMPARISIONS WITH CASE STUDIES

1) EVALUATION ON BENCHMARK CANCER GENE EXPRESSION DATA WITH THREE OBJECTIVES

Further studies with three objectives including the relevance, redundancy, and cardinality of the subsets are conducted to validate the efficiency of our proposed algorithms MOECSA. Under the perspective of different multiobjective algorithms, in this section six efficient multiobjective algorithms, including NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE, are applied to validate the superior performance of MOECSA with three objectives on 35 benchmark datasets.

For the evaluation metrics R_n. Supplementary Table S9 shows the high performance of MOECSA comparing with other multiobjective evolutionary algorithms on 35 cancer gene expression datasets by the measurement R_n. MOECSA performs best among these multiobjective algorithms on 35 benchmark datasets while NSGA-II performs worst. From the statistical results of the last row in Supplementary Table S9, MOECSA performs better than NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE on 32, 27, 33, 33, 22, and 22 datasets respectively. While MOECSA is inferior to other algorithms on 2, 6, 1, 1, 10, and 10 datasets respectively. All the multiobjective evolutionary algorithms can provide the best result "1" on Nutt-2003-v3 maybe because of the small size of samples. The diagnosis performance measured by R_n across 35 benchmark cancer gene expression datasets is depicted in Fig. 5(a) of different multiobjective evolutionary algorithms in boxplot. It is clearly shown that MOECSA can perform better than other compared algorithms generally.

For the evaluation metrics *NMI*, the experimental results are summarized in Supplementary Table S10 on 35 cancer gene expression datasets. For NSGA-II, it can provide the best result "1" on Nutt-2003-v3 and MOECSA can obtain better results on 33 datasets except for Nutt-2003-v2 and Nutt-2003-v3. For MODE, it is inferior to, equal to, superior to our proposed algorithm MOECSA on 27, 6, 2 cancer gene expression datasets respectively. In terms of PESA-II, MOECSA performs better than it on 33 datasets while

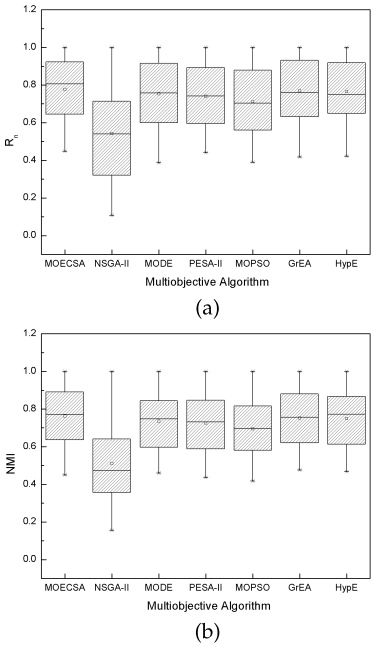


FIGURE 5. Performance of different multiobjective evolutionary algorithms on 35 benchmark datasets with three objectives. The performance is measured by R_n in (a) and *NMI* in (b).

PESA-II outperforms MOECSA on Yeoh-2002-v1. MOPSO also performs better than MOECSA on Yeoh-2002-v1 and it is inferior to MOECSA on 33 datasets. With respect to GrEA, it is superior to MOECSA on 10 cancer gene expression datasets. And it is equal to MOECSA on Alizadeh-2000-v2, Liang-2005, and Nutt-2003-v3 while MOECSA outperforms GrEA on 22 datasets. For HypE, MOECSA can give better results on 22 datasets. There are 10 datasets that are superior to MOECSA showing the effectiveness of HypE. The experimental results of GrEA and HypE show that GrEA and HypE perform better than NSGA-II, MODE, PESA-II, and MOPSO. Besides, Fig. 5(b) exhibits the diagnosis performance across the 35 cancer gene expression datasets measured by *NMI* boxplot of different multiobjective evolutionary algorithms intuitively. It is demonstrated that the overall performance of MOECSA is superior to other compared algorithms.

As can be seen in the experimental results, MOECSA can achieve superior performance among all the compared multiobjective evolutionary algorithms for diagnosing the thirty-five cancer gene expression datasets with three objectives.

2) EVALUATION ON OTHER CANCER GENE EXPRESSION DATA

Four independent disease datasets [36] summarized in Supplementary Table S11 are adopted to validate the performance of MOECSA in further. All cancer gene expression datasets are achieved from (http://portals.broadinstitute.org/cgi-bin/cancer/datasets.cgi). We compare our proposed algorithm with six effective multiobjective evolutionary algorithms that are NSGA-II, MODE, PESA-II, MOPSO, GrEA, and HypE respectively in terms of these four datasets. All algorithms

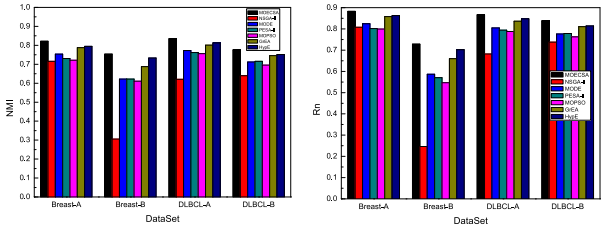


FIGURE 6. **Performance of different multiobjective evolutionary algorithms on four independent disease datasets. The horizontal axis denotes different datasets while the vertical axis denotes *NMI* in (a) and R_n in (b). Different colors indicate different algorithms.**

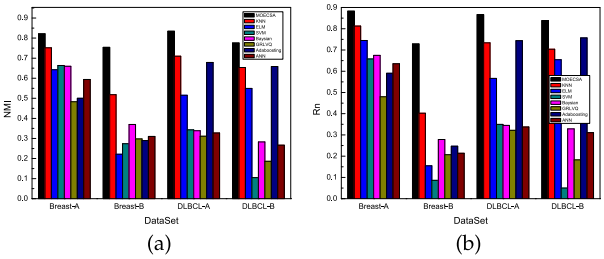


FIGURE 7. **Performance of different classification algorithms on four independent disease datasets. The horizontal axis denotes different datasets while the vertical axis denotes *NMI* in (a) and R_n in (b). Different colors indicate different algorithms.**

carry out 30 independent runs on each dataset. The results of the evaluating indicators *NMI* and R_n for different multiobjective evolutionary algorithms can be depicted clearly in Fig. 6. It can be found that MOECSA performs better than all the other algorithms especially NSGA-II for cancer subtype diagnosis. Meanwhile, the experimental results of MOECSA compared with different classification algorithms, namely KNN, ELM, SVM, Baysian, GRLVQ, Adaboosting, and ANN, evaluated by *NMI* and R_n are shown in Fig. 7. It illustrates MOECSA can give the best solutions among all the classification algorithms. It can be concluded that MOECSA is more reliable to predict the exact label of different genes. In summary, it can be observed that the proposed algorithm MOECSA has high ability to obtain good results for the cancer subtype diagnosis problem with distinct cancer types.

3) EVALUATION ON COLON ADENOCARCINOMA (COAD) DATASET FROM TCGA

We choose the colon adenocarcinoma (COAD) dataset from TCGA (http://tcga-data.nci.nih.gov) to characterize the performance of the proposed algorithm in further. Supplementary Table S15 provides the number of patients and molecular data features of each molecular expression dataset. Each of them has two classes. If the patient suffers from the colon adenocarcinoma, they belong to one class and otherwise if the patient suffers from other cancers, they belong to the other class. In this section, our proposed algorithm is compared with six effective multiobjective algorithms and seven classification algorithms, including NSGA-II, MODE, PESA-II, MOPSO, GrEA, HypE, KNN, ELM, SVM, Baysian, GRLVQ, Adaboosting, and ANN. Four molecular datatypes

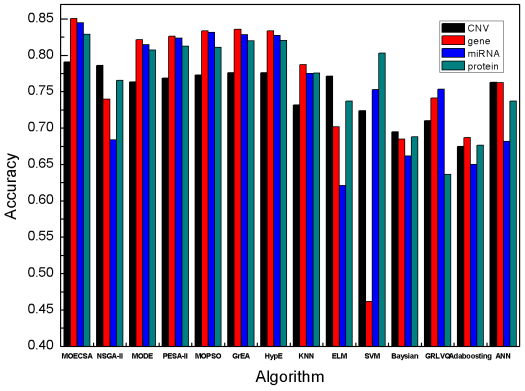


FIGURE 8. **Comparison performance of different algorithms on COAD in terms of accuracy.**

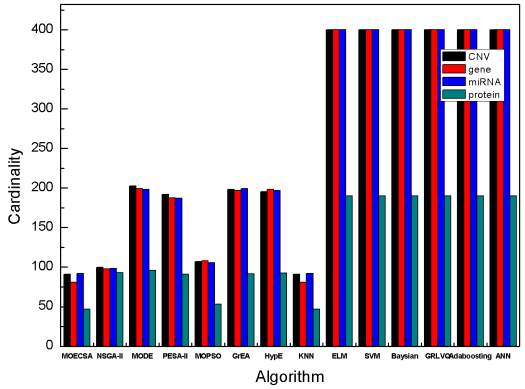


FIGURE 9. **Comparison performance of different algorithms on COAD in terms of cardinality.**

of COAD, including CNV, gene, miRNA, and protein, are used to evaluate the performance of each algorithm. Each algorithm carries out 30 independent runs. The accuracy and the number of features of each algorithm are calculated on four COAD molecular datatypes. They are summarized in Fig. 8 and Fig. 9 respectively. From Fig. 8, MOECSA exhibits a good performance in gaining high accuracy on each datatype of COAD compared with other algorithms. The highest accuracy is gained by MOECSA on the gene datatype of COAD while the lowest accuracy is achieved by SVM on the gene datatype of COAD. The datatype rank of MOECSA with descendant accuracy is the gene, miRNA, protein, and CNV. From Fig. 9, MOECSA is significantly superior to other algorithms in minimizing number of features. The traditional classification algorithms all produce 400, 400, 400, and 190 features on the four molecular datatypes of COAD respectively. For MOECSA, it wins the least number of features on the protein datatype of CODA. For MODE, it achieves the maximum number of features on the gene datatype of COAD. In conclusion, it is clearly shown that MOECSA outperforms other algorithms in predicting the exact label of each molecular datatype on COAD with better accuracy and less number of features. MOECSA can gain good results on sequencing-based dataset for cancer subtype diagnosis problems.

IEEE Journal of Translational Engineering in Health and Medicine

III. CONCLUSION

This paper proposes multiobjective ensemble cuckoo search algorithm, a novel decomposition multiobjective algorithm based on the cuckoo search framework. In order to demonstrate its robust performance, experiments are carried out on thirty-five real cancer gene expression datasets by comparing our proposed algorithm MOECSA with six state-of-the-art multiobjective algorithms and seven effective classification algorithms; the comparisons are based on the evaluation metrics R_n and NMI. In particular, a novel binary encoding is applied to select a small subset of informative genes for multiobjective classification on cancer gene expression data. Four objective functions are defined to capture and interpret multiple characteristics of the cancer gene expression data comprehensively. Finally the cuckoo search algorithm blended with the efficient ensemble mechanism using the decomposition approach optimizes those four objectives simultaneously. Through the pair-wise benchmark comparisons, it is found that MOECSA can obtain competitive performance balancing between convergence and diversity for multiobjective classification on the related gene expression data for cancer subtype diagnosis. In addition, we have created a user-friendly executive software for people to use the proposed algorithm. It is available at https://github.com/wangyh082/MOECSA.git.

REFERENCES

[1] J. Li, H. Su, H. Chen, and B. W. Futscher, "Optimal search-based gene subset selection for gene array cancer classification," *IEEE Trans. Inf. Technol. Biomed.*, vol. 11, no. 4, pp. 398–405, Jul. 2007.

[2] M. S. Mohamad, S. Omatu, M. Yoshioka, and S. Deris, "A two-stage method to select a smaller subset of informative genes for cancer classification," *Int. J. Innov. Comput. Inf. Control*, vol. 5, no. 10, pp. 2959–2968, 2009.

[3] J.-Y. Yeh, "Applying data mining techniques for cancer classification on gene expression data," *Int. J.*, vol. 39, no. 6, pp. 583–602, 2007.

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

[5] L. Ke, Z. Feng, Z. Xu, K. Shang, and Y. Wang, "A multiobjective ACO algorithm for rough feature selection," in *Proc. 2nd Pacific–Asia Conf. Circuits, Commun. Syst.*, Aug. 2010, pp. 207–210.

[6] N. Spolaôr, A. C. Lorena, and H. D. Lee, "Multi-objective genetic algorithm evaluation in feature selection," in *Proc. Int. Conf. Evol. Multi-Criterion Optim.*, 2011, pp. 462–476.

[7] M. Bhattacharyya, L. Feuerbach, T. Bhadra, T. Lengauer, and S. Bandyopadhyay, "MicroRNA transcription start site prediction with multi-objective feature selection," *Stat. Appl. Genet. Mol. Biol.*, vol. 11, no. 1, 2012, Art. no. 6.

[8] X. Li and M. Yin, "Multiobjective binary biogeography based optimization for feature selection using gene expression data," *IEEE Trans. Nanobiosci.*, vol. 12, no. 4, pp. 343–353, Dec. 2013.

[9] X.-S. Yang and S. Deb, "Cuckoo search via Lévy flights," in *Proc. World Congr. Nature Biologically Inspired Comput. (NaBIC)*, 2009, pp. 210–214.

[10] X.-S. Yang and S. Deb, "Multiobjective cuckoo search for design optimization," *Comput. Oper. Res.*, vol. 40, no. 6, pp. 1616–1624, Jun. 2013.

[11] S. Hanoun, S. Nahavandi, D. Creighton, and H. Kull, "Solving a multiobjective job shop scheduling problem using Pareto archived cuckoo search," in *Proc. IEEE 17th Conf. Emerg. Technol. Factory Autom.*, vol. 43, no. 8, 2012, pp. 1–8.

[12] L. S. Coelho, F. A. Guerra, N. J. Batistela, and J. V. Leite, "Multiobjective cuckoo search algorithm based on duffing's oscillator applied to Jiles-Atherton vector hysteresis parameters estimation," *IEEE Trans. Magn.*, vol. 49, no. 5, pp. 1745–1748, May 2013.

[13] A. Syberfeldt, "Multi-objective optimization of a real-world manufacturing process using cuckoo search," in *Cuckoo Search and Firefly Algorithm*. Cham, Switzerland: Springer, 2014, pp. 179–193.

[14] J. Liang and H. K. Kwan, "Fir filter design using multiobjective Cuckoo search algorithm," in *Proc. IEEE 30th Can. Conf. Electr. Comput. Eng.*, Apr./May 2017, pp. 1–4.

[15] G. George and L. Parthiban, "Multi objective fractional cuckoo search for data clustering and its application to medical field," *J. Med. Imag. Health Informat.*, vol. 5, no. 3, pp. 423–434, 2015.

[16] M. Zhang, H. Wang, Z. Cui, and J. Chen, "Hybrid multi-objective cuckoo search with dynamical local search," *Memetic Comput.*, vol. 10, no. 2, pp. 199–208, 2018.

[17] G. Karakaya, S. Galelli, S. D. Ahipaşaoğlu, and R. Taormina, "Identifying (quasi) equally informative subsets in feature selection problems for classification: A max-relevance min-redundancy approach," *IEEE Trans. Cybern.*, vol. 46, no. 6, pp. 1424–1437, Jun. 2016.

[18] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bioinformatics Comput. Biol.*, vol. 3, no. 2, pp. 185–205, 2005.

[19] 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.

[20] 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.

[21] J. Pohjalainen, O. Räsänen, and S. Kadioglu, "Feature selection methods and their combinations in high-dimensional classification of speaker likability, intelligibility and personality traits," *Comput. Speech Lang.*, vol. 29, no. 1, pp. 145–171, 2015.

[22] S. Das, A. Abraham, U. K. Chakraborty, and A. Konar, "Differential evolution using a neighborhood-based mutation operator," *IEEE Trans. Evol. Comput.*, vol. 13, no. 3, pp. 526–553, Jun. 2009.

[23] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE Trans. Evol. Comput.*, vol. 13, no. 2, pp. 398–417, Apr. 2009.

[24] 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.

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

[26] X. Li and K.-C. Wong, "Evolutionary multiobjective clustering and its applications to patient stratification," *IEEE Trans. Cybern.*, to be published.

[27] U. K. Sikdar, A. Ekbal, and S. Saha, "MODE: Multiobjective differential evolution for feature selection and classifier ensemble," *Soft Comput.*, vol. 19, no. 12, pp. 3529–3549, 2015.

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

[29] J. Moore, R. Chapman, and G. Dozier, "Multiobjective particle swarm optimization," in *Proc. Int. Conf. Hybrid Intell. Syst.*, 2000, pp. 56–57.

[30] S. Yang, M. Li, X. Liu, and J. Zheng, "A grid-based evolutionary algorithm for many-objective optimization," *IEEE Trans. Evol. Comput.*, vol. 17, no. 5, pp. 721–736, Oct. 2013.

[31] J. Bader and E. Zitzler, "HypE: An algorithm for fast hypervolume-based many-objective optimization," *Evol. Comput.*, vol. 19, no. 1, pp. 45–76, Mar. 2011.

[32] T. Denoeux, "A k-nearest neighbor classification rule based on Dempster-Shafer theory," *IEEE Trans. Syst., Man., Cybern.*, vol. 25, no. 5, pp. 804–813, May 1995.

[33] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[34] S. Moustakidis, G. Mallinis, N. Koutsias, and J. B. Theocharis, "SVM-based fuzzy decision trees for classification of high spatial resolution remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 50, no. 1, pp. 149–169, Jan. 2012.

[35] B. Xue, M. Zhang, W. N. Browne, and X. Yao, "A survey on evolutionary computation approaches to feature selection," *IEEE Trans. Evol. Comput.*, vol. 20, no. 4, pp. 606–626, Aug. 2016.

[36] Y. Hoshida, J. P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, "Subclass mapping: Identifying common subtypes in independent disease data sets," *PLoS ONE*, vol. 2, no. 11, 2007, Art. no. e1195.