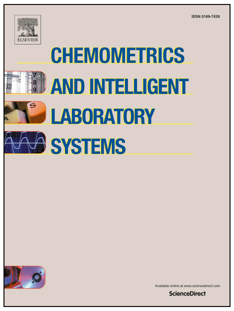
Accepted Manuscript



Hybrid binary Coral Reefs Optimization algorithm with Simulated Annealing for Feature Selection in high-dimensional biomedical datasets

Chaokun Yan, Jingjing Ma, Huimin Luo, Patel Ashutosh

PII: S0169-7439(18)30438-6

DOI: https://doi.org/10.1016/j.chemolab.2018.11.010

Reference: CHEMOM 3712

To appear in: *Chemometrics and Intelligent Laboratory Systems*

Received Date: 6 August 2018

Revised Date: 2 November 2018

Accepted Date: 14 November 2018

Please cite this article as: C. Yan, J. Ma, H. Luo, P. Ashutosh, Hybrid binary Coral Reefs Optimization algorithm with Simulated Annealing for Feature Selection in high-dimensional biomedical datasets, *Chemometrics and Intelligent Laboratory Systems* (2018), doi: https://doi.org/10.1016/j.chemolab.2018.11.010.

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

ACCEPTED MANUSCRIPT

Hybrid Binary Coral Reefs Optimization Algorithm with Simulated Annealing for Feature Selection in High-Dimensional Biomedical Datasets

Chaokun Yan[+a], Jingjing Ma[+a], Huimin Luo[*a], Patel Ashutosh[b]

[a] School of Computer and Information Engineering, Henan University, Kaifeng, 475004, China
[b] VU college, Victoria University, 3011, Melbourne, Australia

Abstract: The last decades have witnessed accumulation in biomedical data. Though they can be analyzed to enhance assessment of at-risk patients and improve the diagnosis, a major challenge associated with biomedical data analysis is the so-called "curse of dimensionality". For the issue, an improved Coral Reefs Optimization algorithm for selecting the best feature subsets has been proposed. Tournament selection strategy is adopted to increase the diversity of initial population individuals. The KNN classifier is used to evaluate the classification accuracy. Experimental results on thirteen public medical datasets show proposed BCROSAT outperforms other state-of-the-art methods.

Keywords: **F**eature selection; Biomedical dataset; Coral Reefs Optimization; Tournament selection; Simulated Annealing

1 Introduction

Large amounts of biological and clinical data have been generated and collected at an unprecedented speed and scale[1]. Such a large amount of data cannot be processed directly by the experts in a short time for diagnosis or treatment, which is giving rise to new requirements for data mining and machine learning. For instance, colorectal microarray dataset include 2,000 features with the highest minimal intensity across 62 samples[2]. These features may cause misleading in the modeling of algorithms for disease diagnoses and overfitting with long training time. In light of the challenges in extracting valuable information and determine the important features of large datasets, feature selection[3][4], also known as variable selection or attribute selection has attracted increasing interests in biomedical domain.

Based on whether the evaluation step includes a learning algorithm or not, feature selection methods can be classified into two categories: filter approaches and wrapper approaches. In filter model, features are evaluated only based on the general data characteristics without utilizing any mining algorithms, which is effective in terms of computational cost. TRank algorithm[5] is one filtering algorithm which has been commonly used in testing the difference of a feature between two groups. However, the main drawback of filter model is that the dependencies among the features are ignored and selected feature subsets could contain some redundant information, which results in worse classification accuracy. In wrapper model, feature subset selection relies on a classifier[6], which can obtain more accurate than the filter approaches because of considering the interactions among the features. Recently, wrapper approach attract increasing attention even though it might require some computational cost for classifier training compared with filter model.

The wrapper approach mainly focused on two aspects of the procedure, namely, subset search and subset evaluation. The former refers to select a subset of features based on corresponding strategy, the latter refers to evaluate the quality of current selected best feature subset and decide whether replace preselected feature subset. For subset search, exhaustive search for the best feature subset of a given dataset would be impossible practically and would encounter the problem of combinatorial explosion[7]. Compared with exhaustive search, Branch and Bound[8] use monotonic evaluation functions and reduce the time cost. However, it's difficult to design evaluation function and can not tackle high-dimensional biomedical problem. In recent years, addressing the problem of feature search through meta-heuristic methods get much attention due to their ability to seek for global optimal solution. The coral reef optimization (CRO)[9-11] algorithm is a novel metaheuristic algorithm that for optimization problems. For example, S. Salcedo-Sanz[12] applied CRO in engineering optimization problems. In the paper, for the high-dimensional biological medical dataset, an improved CRO algorithm has been proposed and show promising behavior in optimizing the feature subset selection.

The main contributions of this paper are summarized as follows:

[+] These authors contributed to the work equally and should be regarded as co-first authors.

* Corresponding authors. Email: luohuimin@csu.edu.cn ckyan@henu.edu.cn

ACCEPTED MANUSCRIPT

- A novel framework based on an improved Coral Reefs Optimization is applied to feature selection for biomedical data.
- Tournament selection strategy is employed to produce initial coral reef populations.
- Simulated Annealing (SA) is combined with CRO to improve the search performance of the original CRO algorithm.

The remainder of this paper is organized as follows. In Section 2, related work on feature selection is presented. In Section 3 the basic principle of CRO algorithm is introduced. The detailed implementation of BCROSAT based feature selection will be explained in Section 4. The experimental results and analysis of the proposed approach are presented in Section 5. Finally, the conclusion summarized in Section 6.

2 Related works

Feature selection is known as an NP-hard and combinatorial problem[13-14]. In recent years, addressing the optimization problem of feature search through meta-heuristic methods get much attention due to their ability to seek for global optimal solution. Many search algorithms have been used to solve feature selection problem on medical datasets, such as Genetic Algorithm (GA), Binary Particle Swarm Optimization (BPSO), etc.

A binary GA[15] was proposed to reduce the number of features, which can simultaneously optimize the parameters and feature subset without degrading the SVM classification accuracy. However, the approach performs slower in terms of convergence speed. Olabiyisi et al.[16] proposed a hybrid GA–SA metaheuristic algorithm for feature extraction, where the GA selection process was combined with SA to avoid getting into the local optima.

Susana et al.[17] presented a modified binary particle swarm optimization (MBPSO) method for mortality prediction of septic patients. Navid et al.[18] proposed a new model based on PSO and Naive Bayesian classification to diagnosethe Parkinson disease. In this model, the optimal training data were selected using PSO algorithm, and then were used to train Naive Bayesian classification.

Azar et al.[19][20] proposed the new supervised feature selection methods based on PSO. The proposed methods combined the advantages of Rough Set Theory (RST) and PSO to solve the medical diagnosis problems. The experimental results show that the hybridization approaches increases the predictive accuracy for biomedical datasets.

A Binary Artificial Bee Colony (BABC) algorithm[19] was used to find the optimal feature subset in the heart disease identification, and then the KNN model was utilized to evaluate these selected features. The performance of BABC has been validated on Cleveland Heart disease dataset. Mafarja et al.[22] proposed an attribute reduction method based on Ant Colony Optimization algorithm, and the rough set theory has been used to evaluate the solutions. However, it is only effective for small datasets, and the process is time-consuming for high-dimensional biological medical dataset.

A novel nature-inspired algorithm named Antlion optimizer (ALO) is proposed by Emery in 2016[23], which has been applied in the feature selection problem. Results show that the proposed BALO algorithm can adaptively search the space of features optimally and obtain better solution.

Hu et al.[24] presented a nature inspired approach Improved Shuffled Frog Leaping Algorithm (ISFLA), which has been successfully applied to feature selection problem in molecular diagnosis of disease. The proposed algorithm has improved the classification accuracy and the efficiency of disease diagnosis by introducing a chaos memory weight factor, an absolute balance group strategy and an adaptive transfer factor.

Majdithe et al.[25] introduced a hybridization model (WOASAT) for feature selection. WOA uses random selection mechanism to select the random solution that enables the algorithm to explore the feature space. SA was adopted to improve the search ability of WOA.The performance of the proposed approach is evaluated on standard benchmark datasets downloaded from UCI repository with a small number of features. However, the biomedical data encountered by researchers is usually high-dimensional, which means having a large number of features. Some existing approaches cannot be suitable effectively for feature selection of biomedical dataset though they performs well in some low-dimensional public datasets. In addition, none of the metaheuristics can solve all optimization problems. Therefore, we need to explore better search strategy to improve feature selection performance.

3 The basic principle of Coral Reef Optimization algorithm (CRO)

As mentioned, the random Search algorithm is an essential part of the wrapper model in feature selection. Recently, some new algorithms are proposed to improve feature selection such as the Coral Reef

ACCEPTED MANUSCRIPT

Optimization algorithm (CRO)[26-27]. In general, the process of the algorithm CRO contains two stages, namely reef formation and coral reproduction, respectively. At the stage of reef formation, the CRO algorithm is first initialized by assigning some squares in the grid. Fig.1 shows the reef model using a 6 ×5 grid, which illustrates an initialization of the reef with corals and coral colonies. Different coral larvae grow in occupied squares. And the other squares are empty, where new corals can freely settle and grow in the future.

In the second stage, CRO searches for optimal solutions through different operations of coral reproduction. Each step of CRO will produce a coral reef larvae, and each larva is labeled with an associated health function $f(R_{ij})$. The reef larvae with larger f value will survive longer than other larvae and vice versa, which means better solutions will be hold in population. The detailed operations are described as follows:

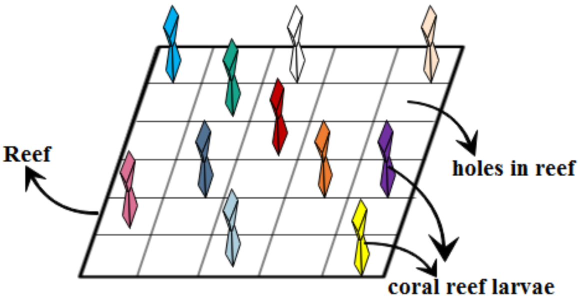


Fig. 1 Coral reef model

Broadcast Spawning (external sexual reproduction): First, some corals are selected with probability F_b to conduct broadcast spawning. Then new larva will be created. With a probability p_k, couples can be selected from the pool of broadcast spawner corals in step *k*. Once two corals have been selected to be the parents of a larva, they are not chosen any more in step *k*.

Brooding (internal sexual reproduction): At each step *k* of the reef formation phase in the CRO algorithm, brooding operation will be achieved by selecting corals with $1-F_b$. The brooding operation consists of a random mutation of the brooding-reproductive coral. The produced larvae is then released out to the water in a similar fashion than that of the larvae generated in step Broadcast.

Larvae setting: Once all the larvae are formed at step *k* by broadcast spawning or brooding, they will try to set and grow in the reef. First, the healthiness value of each coral will be calculated. Then, each larvae will randomly try to set in a square (*i*, *j*) of the reef. If the square is empty, the coral grows there in no matter the value of its health function. By contrast, if a coral is already occupied by a square, the new larvae will set only if its health function is better than that of the existing coral. All new larvae and corals will compete for the space, and then the one with a better fitness value will occupy a grid of reef. Only *k* opportunities for larval settlement, otherwise the larva will be depredated by other animals in the reef.

Budding (asexual reproduction): The overall set of existing corals in the reef are sorted as a function of their level of healthiness. Some corals are duplicated with the fraction F_a and tries to settle in a different part of the reef by following the setting process described in Larvae setting.

Depredation: At the end of each reproduction step *k*, a small number of corals may die during the reefs formation phase. The depredation operator is conducted with a very small probability F_d of the worse health corals in *REEF* , the value of this fraction may be set to $F_a = F_b$.

4 BCROSAT for Feature Selection

In this section, the encoding scheme for feature selection is described firstly. Here, to emphasize the binary encoding strategy for feature selection problem, the BCRO (Binary Coral Reef Optimization) is taken in the following descriptions instead of using CRO. Then we define the evaluation function of the proposed algorithm. Finally, the implementing steps of the proposed BCROSAT for feature selection are described. The flowchart of the BCROSAT for feature selection is shown in Fig. 2.

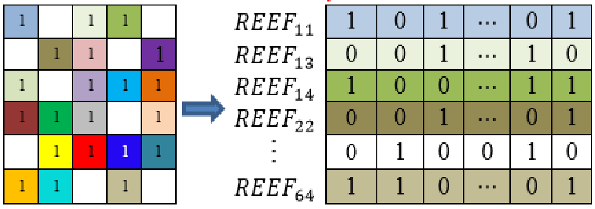
4.1 Encoding

The coral reefs *REEF* consisting of a $N\times M$ square grid is presented in Fig.3(a). The square occupied by coral larvae is set to 1 and the empty is 0. In Fig.3(a), each occupied square $REEF_{ij}$ in *REEF* represents a coral larvae of the coral population *REEFpob*, denoted as $REEFpob_m$. Fig.3(b) shows the initial population *REEFpob* containing $N\times M\times r_0$ larvae, here r_0 denotes a rate value between occupied/total elements, which is used to determines the initial population density (with $0<r_0<1$). Each coral larva is modeled as a binary one-dimensional vector of length *L*. The vector consisting of a series of binary values of 0 and 1 represents a feature subset. For example, the occupied square $REEF_{13}$**Error! Reference source not found.**in Fig.3(a) represents the second row of *REEFpob***Error! Reference source not found.**in Fig.3(b). Different

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

color squares in Fig.3(a) represent different coral larvae of **Error! Reference source not found.**.



(a) REEF (b) REEFpob

Fig. 3 Coral population initialize the encoding

The binary individual *REEFpob*$_m$ represents a candidate solution (a feature subset), *m* is the number of squares occupied by coral larvae, where $m\in N\times M\times r_0$. Each coral larva *REEFpob*$_m$ has various features (f_1, f_2, ... , f_j ($j\in$ L)), the value of each feature f_j is set to 1 if the *j*th feature is selected, otherwise 0. The parameter *pro* represents the probability of selecting the relevant combination of features, *r* is a random number in {0,1} the feature f_j is selected randomly in the following Eq.(1).

$$REEFpob_{ij} = \begin{cases} 1, r < pro \\ 0, r \geq pro \end{cases} j = 1, 2, 3 \ldots L \quad (1)$$

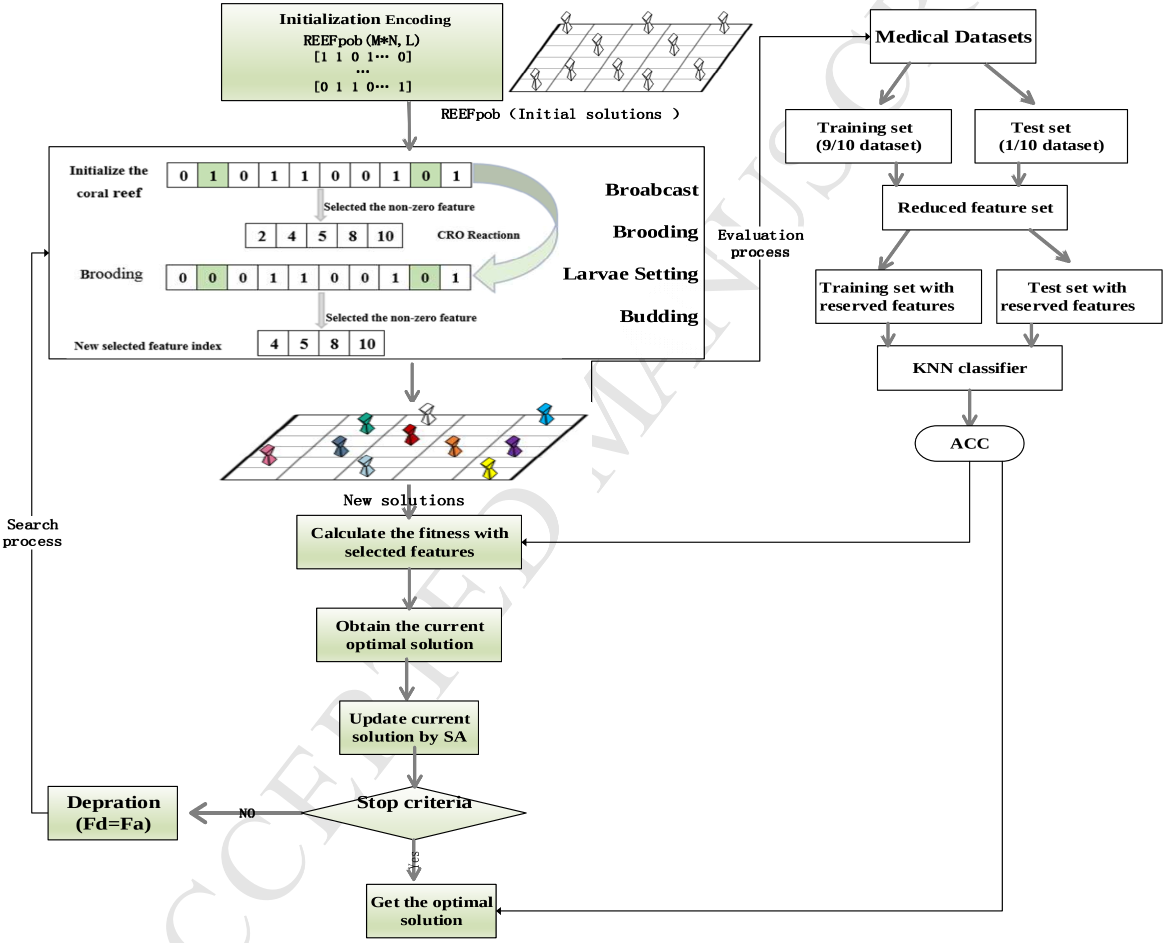


Fig. 2 The flowchart of BCROSAT for feature selection

4.2 Fitness function

Feature selection involves two main objectives: maximize the classification accuracy and minimize the number of features. In this study, we use the proposed meta-heuristic method BCROSAT to select the optimal feature subset, and use the KNN-based fitness function defined in Ref.[25] for evaluation.

$$acc(KNN) = \frac{num_c}{num_c + num_i} \times 100\% \quad (2)$$

$$fitness = \omega_1 \times acc(KNN) + \omega_2 \times (1 - \frac{n}{N}) \quad (3)$$

Here, **Error! Reference source not found.**and **Error! Reference source not found.**are set to 1 and 0.001, respectively, as in Ref.[24]. The function *acc(KNN)* refers to classification accuracy based on KNN. *N* is the total number of features and *n* is the

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

number of selected features. The numbers of correctly and incorrectly classified examples are indicated by **Error! Reference source not found.** and **Error! Reference source not found.** respectively.

4.3 Implementation steps of proposed BCROSAT

Each coral larva represents a solution, which is settled by competing for growth space. The new solution continuously updated by coral reproduction mentioned above. Here, SA is adopted to escape local optima by allowing hill-climbing moves in hopes of finding a global optimum. The approach BCROSAT combines the advantages of global search algorithm BCRO and local search of SA[25] to solve feature selection problems. The detailed implementation of proposed algorithm is described as follows.

Step1: Initialization: Initialize the coral reef model *REEF* and coral reef population *REEFpob.*

Randomly assign some grids in *REEF* to construct coral populations. As mentioned as section 4.1, we model a solution by binary string. Here, the Tournament selection mechanism (TS)[28] is used to update the worst solution and generate an new initial population. Fig.4 show the process of population initalization. We choose *m* solutions randomly from the model *REEF*, these solutions are compared against each other and a tournament strategy will be applied to determine the winner. The tournament involves a random number *r* in range [0,1], the selection probability is set to 0.5. The solution with the highest fitness value will be selected while the value of *r* is greater than 0.5, otherwise the weak solution will be chosen. During the initialization process of BCROSAT, the individual selected by TS replaces the worst individual in the coral reef population.

Step 2: Search mechanism: A best feature subset with the maximum health function value is selected.

We implement broadcast spawning process by conducting crossover operator. Two corals P_1 and P_2 will be selected randomly and produce the offspring L_1 and L_2 by external sexual reproduction, which can be regarded as a two-point crossover operator of GA[29]. Fig.4 presents how to broadcast spawning of BCRO. In this example, the blue part of P_1 will be exchanged with the yellow part of P_2, and then these new results will be their offspring.

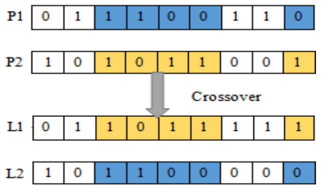


Fig. 4 Example of broadcast spawning

The brooding process is implemented by random mutations of coral larvae shown in Fig.5. First we randomly choose one coral larvae *P* and invert it to a new solution *L* or randomly choose two subsolutions and exchange them. Then neighborhood search is used to find a better solution. After broadcast spawning, the larvae can improve the fitness value of each solution.

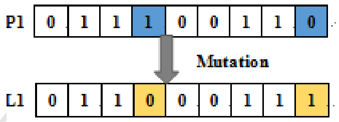


Fig. 5 Example of Brooding

For larvae setting, the best larvae will be settled in the reef, while the larvae with the worst fitness value will be eliminated. The new solution produced in sexual reproduction search a grid to settle down. Only the healthier corals with greatest fitness value can settle successfully.

In the asexual stage, all corals in the reef are sorted according to the values of health function. Then, the better corals will be duplicated to produce new solutions. Finally, a fraction of the worse solutions is eliminated in depredation.

Step 3: Fitness Evaluation: To evaluate the quality of each solution, Eq.(3) is utilized to obtain the fitness value of each coral in the reef, which represents the level of healthiness of corals.

Step 4: Update solution: The best solution is updated by SA with strong local search ability to search the neighbors of the optimal solution.

Compared with other optimization algorithms, SA can be considered as a hill-climbing based method that iteratively try to improve a candidate solution with respect to the objective function. A better neighbor solution will be accepted, whilst the worse neighbor is selected with a certain probability to help the algorithm escape from the local optima. The probability of selecting a worse solution is determined by the Boltzmann probability *p.* At each iteration, the algorithm searches the neighborhood of the best solution to obtain a new one with an improved functional value. After the corals' reproduction is completed, the optimum solution *bestSol* in the population is obtained, and then SA is employed to search its neighborhoods and update *bestSol* by a

ACCEPTED MANUSCRIPT

larvae with better fitness value.

$$p = e^{-\theta / T} \quad (4)$$

Where θ is the difference of the evaluation of the objective function between the best solution and the generated neighbor, the parameter T denotes the temperature, and decreases periodically during the search process according to some cooling schedule.

.**Step 5:** Continue the iterative process: return to Step 2 if stopping criteria have not been met.

The pseudocode of BCROSAT for feature selection is shown as follows:

Input: Problem-specific information (the objective function *f*, *REEF M*N*, the dimension of the *REEFpob L*)
Output: the optimal solution of coral reef

```
Assign parameter values to REEF, REEFpob, T0 and T.
/* Initialization */
Initialize coral reef REEF and REEFpob by Tournament selection
Calculate the fitness value of each coral by Eq. (3);
/* The search process of coral reef algorithm */
while (the stopping criteria not met) do
 /* operation1: Sexual reproduction */
    Sexual crossover process by broadcast spawning;
    for couples of broadcast spawning corals $R_{ij}$ and $R_{kj}$ do
        $R_{ij} + R_{kj} \rightarrow R^0 + R^1$
    end for
    Sexual reproduction by brooding;
    for each brooding coral $R_{ij}$ do
        $R_{ij} \rightarrow R^2$
    end for
/* operation2: Larvae Setting */
    Settlement of new larvae (competition among species)
/* operation3：Asexual Reproduction */
    Asexual Reproduction budding;
    Duplicate the better corals in the reef to produce new solutions
/*operation4：Depredation */
    Coral depredation
    Evaluate each coral larvae by KNN classifier and compute fitness
    Obtain the current optimal solution bestSol with maximum fitness value
/*operation5：Update current solution by SA */
    while (itr < MaxIt ) do
        $S_i \leftarrow bestsol$
        The fitness value $f(S_i)=f(bestsol)$
        for T=T0
            Generate at random a new solution new in the neighbor of $S_i$ calculate $f(S_i)$;
            if $f(S_i) > f(new)$
                $S_i \leftarrow new; bestsol \leftarrow new;$
                $f(S_i) \leftarrow f(new)$
                $f(bestsol) \leftarrow f(new)$
            else
                Calculate $\theta = f(new)- f(bestsol)$
                Generate a random number, $P = [0,1]$;
                if ($P < e^{-\theta / T}$ )
                    $S_i \leftarrow new; f(S_i) \leftarrow f(new)$
                end if
            end if
            Update temperature $T= alpha*T$
        end for
    end while
end while
Get the optimal solution g
```

Algorithm.1 BCROSAT Algorithm for Feature Selection

Where *REEFpob* represents the population size of the coral reef, *P1* and *P2* represent the new individual produced by the broadcast spawning. *P* represents a new individual with a brooding process. *MaxIt* is the number of iterations of SA. *T0* is initial temperature, S_i decides the struct of the neighbors of *bestSol* neighbors, *f*(R) is used to calculate corals' level of healthiness.

5 Experimental results and analysis

The algorithm proposed has been implemented on a PC with Intel Dual Core CPU, 4 GB RAM, and Windows 7 operating system. In order to evaluatethe performance of BCROSAT method, we compare it with the other four state-of-the-art methods: IGA[15], MBPSO[17], ISFLA[24], WOASAT[25], As the improved version of the classic algorithms GA and PSO, IGA and MBPSO are widely used as feature selections problem. ISFLA explores the space of possible subsets to obtain the set of features that maximize the predictive accuracy and minimizes irrelevant features in high-dimensional biomedical data. The hybridization model WOASAT is a novel meta-heuristic algorithm for feature selection. Besides, in order to verify the effectiveness of our improvement, BCRO is also taken as a benchmark algorithm. The evaluation indicators are described as follows.

ACC%:The average classification accuracy;

AvgN:The average number of feature subsets；

AvgD%: The Average dimension;

AvgF%: The Average fitness;

5.1 Datasets

To evaluate the effectiveness of proposed method, thirteen typical high-dimensional biomedical datasets are utilized in our experiments and listed in Table 1. The dimensional scopes of datasets range from 2000 to 12600. The first nine benchmark datasets can be downloaded from the Kent Ridge Biomedical Dataset

ACCEPTED MANUSCRIPT

at website: http://leo.ugr.es/elvira/DBCRepository/, and they are usually two-class classification datasets, such as ALL-AML-Leukemia, ColonTumor, Nervous-System et al, which provides data with respect to gene expression, protein profiling, genomic sequence and disease diagnosis including. In addition, the rest of the datasets are microarray datasets[31], which are multi-class data and can be downloaded from the http://csse.szu.edu.cn/staff/zhuzx/Datasets.html.

Table.1 Benchmark datasets

Dataset	Instances	Attributes	Classes
ALL-AML_train	38	7130	2
ColonTumor	62	2000	2
DLBCLOutcome	58	7129	2
DLBCL-Stanford	47	4026	2
lungCancer_train	32	12534	2
LungCancerOntario	39	2880	2
NervousSystem	60	7129	2
DLBCL-NIH-train	160	7400	2
lung-harvard1	203	12600	5
MLL	72	12582	3
SRBCT	83	2308	4
Leukemia_3c	72	7129	3
Leukemia_4c	72	7129	4

5.2 Parameter Setting

The values of the BCRO and BCROSAT parameters for all the datasets are initialized as shown in Table 2. We used KNN classifier with 10-fold-CV as a fitness function and feature subset evaluator on these datasets. In the process of 10-fold-CV, the dataset is randomly divided into ten parts averagely, each part is taken as test set in turn with the remaining nine parts as train set. Accuracy and the number of selected features are used to evaluate the performance of these methods. For the fairness of the experiment, the experiments of all methods are also repeated 10 times, and we take the mean values as the final results. In our experiments, the size of the coral reef population and the iteration number are set as 30 and 150, respectively. Other parameter settings are from references[9] [25].

Table.2 Parameter Settings

Parameters	BCRO	BCROSAT
REEFpob	**30**	**30**
iterations	**100**	**100**
Fb	**0.9**	**0.9**
Fa	**0.001**	**0.001**
Fd	**0.01**	**0.01**
r0	**0.7**	**0.7**
k	**3**	**3**
Pd	**0.1**	**0.1**
pro	**_**	**0.5**
T0	**_**	**0.1**
Alpha	**_**	**0.99**
MaxIt	**_**	**30**

5.3 Experimental results and analysis

In this section, we present the experimental results and analysis for all comparative indicators.

5.3.1 Performance comparison

The comparison result in tems of the classification accuracy and the average number of feature subsets is shown in Table 3. As we can see, BCROSAT outperforms all approaches on most datasets. For example, BCROSAT can obtain the highest accuracy and selects the smallest number of features for most datasets including ALL-AML_train, ColonTumor, DLBCLOutcome, lungCancer_train, NervousSystem, lungcancer-harvard1, MLL, Leukemia_3c and Leukemia_4c. For dataset LungCancerOntario, BCROSAT provides 86.75% accuracy by using about 13 attributes, whilst the number of features is slightly worse than that of MPSO. For DLBCL-NIH-train, ISFLA can obtain a higher accuracy but select more features than the proposed method BCROSAT. Comparatively speaking, WOASAT and MPSO obtain worse accuracy compared with other algorithms on almost all datasets. Except for DLBCL-Stanford, LungCancerOntario, and SRBCT, BCROSAT can obtain the smallest number of feature (AvgN) on other nine datasets shown in Table4. There is no conclusive evidence that the fitness value will increase with using more features. For example, the AvgN obtained by WOASAT is 46.2, 61.9 and 50 for DLBCLOutcome, Leukemia_3 and Leukemia_4 respectively, which is worse than other algorithms, while WOASAT do not perform better with respect to classification accuracy.

The Fig.6 presents the comparison for average dimension, which reflects the percentage the number of feature subset selected to maximum dimension of the original feature set. For ColonTumor and SRBCT dataset, IGA and MPSO obtain more feature number than other algorithms. It is significantly observed that BCROSAT has the lower average dimension except for DLBCL-Stanford, LungCancerOntario and SRBCT. A smaller feature number meas less feature redundant, which further verify the good performance of proposed BCROSAT.

As shown from Fig.7, the proposed algorithm outperforms others for average fitness value on high-dimensional biomedical datasets. For datasets ALL-

ACCEPTED MANUSCRIPT

AML_train, and lungCancer_ train, the average fitness obtained by the BCROSAT are almost 100%, which are very close to the result obtained by other algorithms. For other benchmark datasets, the fitness values achieved by BCROSAT is greater than that obtained by WOASAT, IGA, and MPSO, especially for LungCancerOntario and Leukemia_4c dataset. The experimental results further show proposed BCROSAT algorithm can perform well for feature selection of biomedical datasets.

Table.3 The experiment results of five algorithms on the benchmark datasets

Datasets	ACC(%)					AVGN				
	BCROSAT	WOASAT	ISFLA	IGA	MPSO	BCROSAT	WOASAT	ISFLA	IGA	MPSO
ALL-AML_train	**99.25**	92.5	95.84	92.43	91.71	**18.5**	30.9	35.32	34.75	38.75
ColonTumor	**92.31**	82.14	88.89	85.62	82.17	**20.5**	20.9	35.83	38.33	42.16
DLBCLOutcome	**77.49**	70.67	72.87	68.12	62.19	**23.16**	46.2	26.37	30.13	27.25
DLBCL-Stanford	**97.20**	86.00	92.40	86.30	83.24	17.6	26.1	**15.125**	16.75	16.62
lungCancer_train	**100.0**	96.07	99.17	97.48	95.85	**29.74**	40.7	50	48.38	47.38
LungCancerOntario	**86.75**	78.17	83.12	79.32	76.06	13.3	16.70	15.00	11.88	**10.05**
NervousSystem	**82.00**	70.00	75.06	70.18	68.87	**21.4**	45.2	30.38	29.125	25.27
DLBCL-NIH-train	70.12	68.75	**71.47**	65.09	59.66	**18.5**	32	29.88	28.75	30.50
lungcancer-harvard1	**93.57**	90.11	93.11	90.65	89.35	**23.3**	33.3	25.0	29.66	25.33
MLL	**98.04**	85.64	92.62	89.67	90.64	**35.6**	55.80	55.7	63.88	65.77
SRBCT	**95.76**	77.36	80.24	89.97	85.17	33	**20.20**	39.08	39.33	47.55
Leukemia_3c	**94.50**	83.04	90.91	85.43	83.45	**32**	61.90	38.65	56.88	34.44
Leukemia_4c	**90.90**	76.71	84.12	81.17	73.68	**30.9**	50	37.55	31.77	36.11

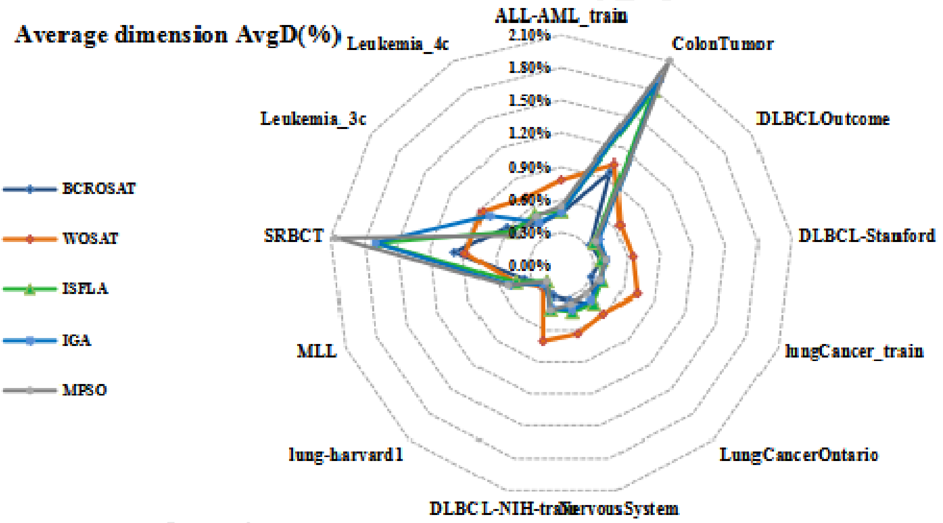


Fig.6 Average dimension comparison

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

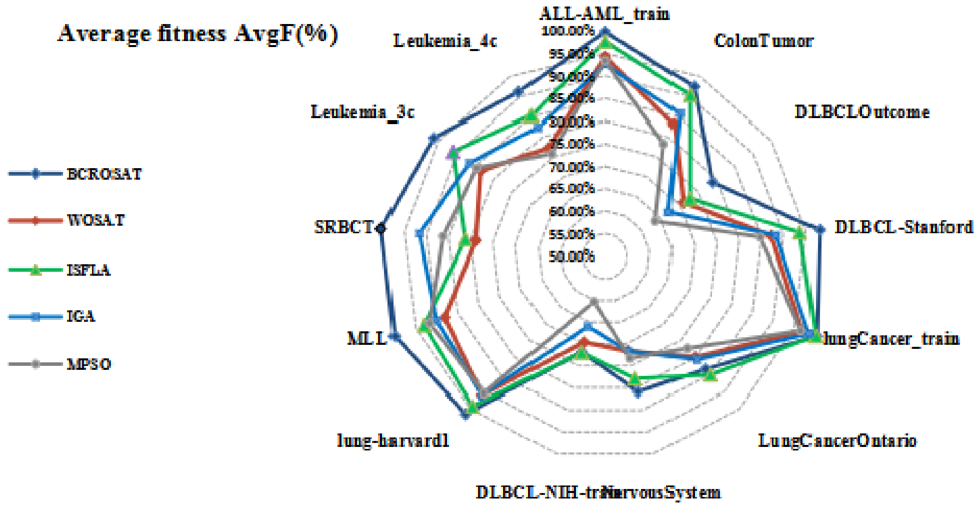


Fig.7 The average fitness comparison

The aim of feature selection is to reduce the dimensionality of the original data and to improve the efficiency of search mechanism. The time consumption for feature selection of high-dimensional biological datasets is also considered here. Fig.8 presents the average computational time comparison for all algorithms. As can be seen from the results, time cost of BCROSAT achieves better performance on most benchmark disease datasets except for lungCancer_train, NervousSystem and lung-Harvard1. Basically, the running time of these algorithms depend on both their convergence speed and the scale of datasets. For example, the running time on lung-harvard1 is more than MLL. The reason is that the dataset lung-harvard1 has more samples and more categories. In general, compared with IGA, MPSO, ISFLA, and WOASAT, proposed BCROSAT is more efficient in terms of time cost.

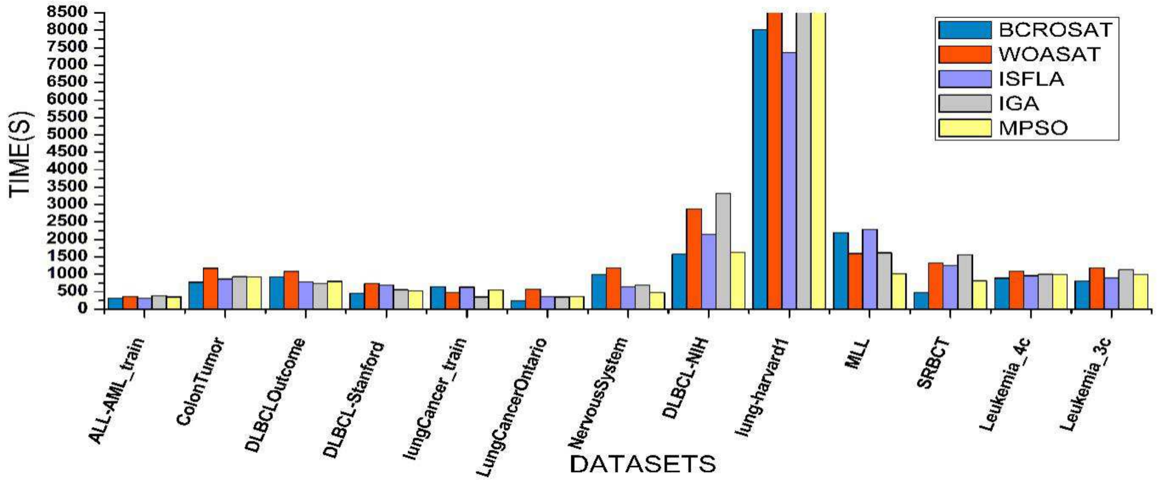


Fig.8 Running time comparison

Table.4 The experiment results obtained byBCROSAT and BCRO

Datasets	ACC(%)		AVGN	
	BCROSAT	BCRO	BCROSAT	BCRO
ALL-AML_train	**99.25**	97.49	**18.5**	36.26
ColonTumor	**92.31**	91.47	**20.5**	21.18
DLBCLOutcome	77.49	**77.91**	**23.16**	25.59
DLBCL-Stanford	**97.20**	96.44	17.6	**15.6**
lungCancer_train	**100.0**	98.9	**29.74**	36.42

ACCEPTED MANUSCRIPT

LungCancerOntario	**86.75**	84.75	**13.3**	15.83
NervousSystem	**82.00**	80.58	**21.4**	26.7
DLBCL-NIH-train	**70.12**	65.60	18.5	**17.13**
Lung-Harvard1	93.57	**94.05**	**23.3**	34.2
MLL	**98.04**	97.54	**35.6**	51.7
SRBCT	**95.76**	93.80	33	**27.9**
Leukemia_3c	94.50	**95.69**	**32**	48.50
Leukemia_4c	**90.90**	90.58	**30.9**	38.65

5.3.2 Comparison with the original CRO

To further evaluate the impact of Tournament selection mechanism and SA, we compare BCROSAT with BCRO. As shown in Table 4, the hybrid algorithm BCROSAT is much better than original BCRO in terms of classification accuracy and number of selected feature. For dataset DLBCLOutcome and Leukemia_4c, the ACC of BCROSAT is 77.49% and 90.90%, respectively. The results are close to the ACC obtained by the original BCRO. Meanwhile, the number of selected features is competitive on some datasets. For dataset DLBCL-NIH-train, the original algorithm BCRO gets accuracy 65.60% by using about 17 features. However, BCROSAT obtains accuracy 70.12% by using about 19 features. Likewise, for the dataset DLBCL-Stanford, the proposed algorithm BCROSAT selects 2 more features than the BCRO but achieves higher accuracy.

In general, BCROSAT shows superior performance in terms of classification accuracy and the average number of features for most datasets reported in Table 4. BCROSAT achieved better performance using fewer features, which demonstrates that SA can improve the local search capability of BCRO. Comparatively speaking, BCROSAT was slightly worse than BCRO algorithm in terms of execution time because of the additional search operation caused by SA.

5.3.3 Complexity analysis

The computational time analysis of meta-heuristic algorithm includes initialized preprocessing time and searching time. The population initialization contains two phases: binary coding and fitness evaluation. For encoding into a binary string, the time complexity is O(n). For fitness evaluation, the time complexity is O(n^2)，where n is the number of coral individuals.

The basic operations performed in the search process are broadcast spawning, brooding, larvae setting, budding and local search (SA). The broadcast spawning, brooding have time complexity of O(L^2) and O(L), respectively. Here L is larval individual length. The time complexity of larvae setting is O(n^2). For budding, the operation has a time complexity of O(n). The time complexity of SA is O(n^2). The search process of BCROSAT is repeated till the stopping criteria is met, the time complexity G is given by Eq(5), where $n1$ denotes the number of iterations of convergence.

The total time complexity T is given by Eq(6).

$$G=O(n1\times(L^2+L+n^2+n+n^2)) \tag{5}$$

$$T=O(n+n^2+n1\times(L^2+L+n^2+n+n^2)) \tag{6}$$

5.3.4 Evaluate the impact of the three classifiers

As mentioned above, BCROSAT achieved good classification performance for disease diagnosis. For proposed BCROSAT, KNN classifier was used to evaluate the feature subsets selected. To evaluate the impact of the classifier on the performance of BCROSAT, another two popular classifiers, SVM and Extreme Learning Machine (EML) are tested in terms of accuracy and the number of selected features based on the thirteen benchmark datasets. Similarly, the 10-fold cross validation experiments were performed to evaluate the classifier model, and the results are reported in Table 5. It can be seen that the results based on KNN and SVM classifiers are very close for eight datasets including ALL_AML train, ColonTumor, SRBCT, Leukemia_4c, lungCancer train, LungCancer-Ontario, NervousSystem, DLBCL-NIH-train. Comparatively speaking, EML classification is worse than KNN and SVM, especially for datasets Lung-Harvard1, Leukemia_4c, MLL, Leukemia_3c and SRBCT. For the method with KNN, the standard deviations with respect to ACC and AvgN are less than others in almost all datasets except for the datasets ALL-AML, NervousSystem, and Lung-Harvard1. The smaller the standard deviation, the more stable the experimental results. According to the results, we can conclude that KNN-based BCROSAT has better robustness for feature selection. As can be seen from Table 5, the average accuracy and the average number of feature subsets obtained by BCROSAT with KNN

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

outperformed those obtained by BCROSAT with the other two classifiers.

Table.5 The performance evaluation results of three different classifiers combined with BCROSAT

Dataset	**Classifier**	**Acc(%)**		**AvgN**	
		Acc(%)	**Std**	**AvgN**	**std**
ALL-AML_train	**KNN**	99.25	1.35	18.5	5.97
	SVM	98.75	**1.16**	23.10	6.72
	EML	95.25	3.05	23.60	6.57
ColonTumor	**KNN**	92.31	1.67	20.50	6.84
	SVM	90.86	1.90	21.8	8.37
	EML	87.31	2.49	24.20	9.84
DLBCLOutcome	**KNN**	77.49	1.70	23.16	4.85
	SVM	75.33	3.07	36	5.67
	EML	70.27	2.33	30.6	5.78
DLBCL-Stanford	**KNN**	97.20	1.94	17.6	3.62
	SVM	94.45	2.27	29.7	4.50
	EML	91.95	2.23	24.9	5.13
lungCancer_train	**KNN**	100	0.06	29.74	5.35
	SVM	99.42	1.18	32.6	10.89
	EML	96.17	3.01	42.3	21.49
LungCancerOntario	KNN	86.75	1.59	13.33	6.55
	SVM	85.17	2.29	22.6	7.89
	EML	86.67	2.81	27.7	10.48
NervousSystem	**KNN**	82.00	2.41	21.4	3.87
	SVM	78.17	**2.17**	36.4	9.15
	EML	74.83	1.57	25.6	12.82
DLBCL-NIH-train	**KNN**	70.12	**1.29**	**18.5**	**6.95**
	SVM	69.63	2.21	21.25	7.01
	EML	66.38	1.55	26.6	12.92
Lung-Harvard1	**KNN**	93.57	1.54	30.5	**6.38**
	SVM	95.04	**0.84**	50.2	5.77
	EML	80.26	1.44	**54**	10.71
MLL	**KNN**	**98.04**	**3.58**	**35.6**	**6.37**
	SVM	96.73	1.07	58.3	10.15
	EML	79.70	4.99	50.2	8.24
SRBCT	**KNN**	95.76	**3.08**	33	**5.98**
	SVM	92.08	4.10	35.3	7.81
	EML	82.93	5.13	**32.5**	7.21
Leukemia_3c	**KNN**	**94.50**	**2.21**	**32**	**6.72**
	SVM	91.71	2.26	42.5	7.76
	EML	83.36	3.27	35.2	6.01
Leukemia_4c	**KNN**	**90.90**	**2.37**	**30.9**	**3.39**
	SVM	89.80	4.32	30.85	5.67

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

EML	78.23	3.19	34.4	8.04

6 Conclusion

Feature subset selection is a fundamental technique in many application areas and different evolutionary algorithms have been developed to solve different feature subset selection problems. However, the recent increase of dimensionality of the used data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. Here, a new hybrid algorithm BCROSAT was proposed. Tournament selection mechanism and SA algorithm are combined with BCRO to design the hybrid algorithm BCROSAT which can solve feature selection problem for high-dimensional biomedical data efficiently. KNN classifier serves as an evaluator of our proposed algorithms. Experiment results show that our proposed method can reduce the number of dataset features and achieve a higher classification accuracy simultaneously. For most biological datasets, BCROSAT-KNN can achieve high performance with least number of features in short time when compared with the other art-of-state methods. The proposed method can serve as an ideal pre-processing tool to help optimize the feature selection process of high-dimensional biomedical data, better mine the function of biological datasets in fields of disease diagnosis, and improve the efficiency of disease diagnosis. In the future, we will further improve the exploration and exploitation of BCRO by integrating with other local search strategies or swarm intelligent algorithms.

Acknowledgment

This work was supported in part by the National Natural Science Foundation of China (No. 61802113, No.61802114).

REFERENCES

[1] Lee K, Man Z, Wang D, et al. Classification of microarray datasets using finite impulse response extreme learning machine for cancer diagnosis[J]. Neural Computing & Applications, 2013, 22(3-4):457-468.

[2] . Hira Z M, Gillies D F. A Review of Feature Selection and Feature Extraction Methods Applied on Microarray Data[J]. Advances in Bioinformatics, 2015, 2015:1-13.

[3] H. Liu and Z. Zhao, Manipulating data and dimension reduction methods: Feature selection, in Encyclopedia of Complexity and Systems Science. Berlin, Germany: Springer, 2009, pp. 5348–5359.

[4] Kira K, Rendell L A. The feature selection problem: traditional methods and a new algorithm[C]// Tenth National Conference on Artificial Intelligence. AAAI Press, 1992:129-134.

[5] Baldi P, Long A D. A Bayesian framework for the analysis of microarray expression data: regularized t-test and statistical inferences of gene changes[J]. Bioinformatics, 2001, 17(6): 509-519.

[6] Verbiest N, Derrac J, Cornelis C, et al. Evolutionary wrapper approaches for training set selection as preprocessing mechanism for support vector machines: Experimental evaluation and support vector analysis[J]. Applied Soft Computing, 2016, 38(C):10-22.

[7] Williams R. Improving exhaustive search implies superpolynomial lower bounds[J]. Siam Journal on Computing, 2013, 42(3):231-240.

[8] Yagiura M, Ibaraki T. On metaheuristic algorithms for combinatorial optimization problems[J]. Systems & Computers in Japan, 2015, 32(3):33-55.

[9] Salcedo-Sanz S, Del S J, Landa-Torres I, et al. The coral reefs optimization algorithm: a novel metaheuristic for efficiently solving optimization problems[J]. Thescientificworldjournal, 2014, 2014(8):739768.

[10] Salcedo-Sanz, Sancho, et al. "A novel coral reefs optimization algorithm for multi-objective problems." International Conference on Intelligent Data Engineering and Automated Learning. Springer, Berlin, Heidelberg, 2013.

[11] Yang Z, Zhang T, Zhang D. A novel algorithm with differential evolution and coral reef optimization for extreme learning machine training[J]. Cognitive Neurodynamics, 2016, 10(1):73.

[12] Salcedo-Sanz S. A review on the coral reefs optimization algorithm: new development lines and current applications[J]. Progress in Artificial Intelligence, 2017, 6(1): 1-15.

[13] Salcedo-Sanz S, Pastor-Sánchez A, Prieto L, et al. Feature selection in wind speed prediction systems based on a hybrid coral reefs optimization–Extreme learning machine approach[J]. Energy Conversion and Management, 2014, 87: 10-18.

[14] Woeginger G J. Exact algorithms for NP-hard problems: a survey[C]// Combinatorial Optimization - Eureka, You Shrink!, Papers Dedicated To Jack Edmonds, International Workshop, Aussois, France, March 5-9, 2001, Revised Papers. DBLP, 2003:185-208.

[15] Babatunde O, Armstrong L, Leng J, et al. A genetic algorithm-based feature selection[J]. British Journal of Mathematics & Computer Science, 2014, 4(21): 889-905. Huang, C. L., & Wang, C. J. (2006). A GA-based feature selection and parameters optimizationfor support vector machines. Expert Systems with applications, 31(2), 231-240.

[16] Olabiyisi S O, Fagbola T M, Omidiora E O, et al. Hybrid MetaHeuristic Feature Extraction Technique for Solving Timetabling Problem[J]. International Journal of Scientific & Engineering Research, 2012, 3(8).

[17] Vieira S M, Mendonça L F, Farinha G J, et al. Modified binary PSO for feature selection using SVM applied to mortality prediction of septic patients[J]. Applied Soft Computing, 2013, 13(8):3494-3504.

[18] Ghanad N K, Ahmadi S. Combination of PSO algorithm and Naive Bayesian classification for Parkinson disease diagnosis[J]. Advances in Computer Science: an International Journal, 2015, 4(4): 119-125.

[19] Inbarani H H, Azar A T, Jothi G. Supervised hybrid feature selection based on PSO and rough sets for medical diagnosis.[J]. Computer Methods & Programs in Biomedicine, 2014, 113(1):175.

[20] Velayutham, Thangavel. Unsupervised Quick Reduct Algorithm Using Rough Set Theory[J]. Journal of Electronic Science and Technology, 2011, 9(3):193-201.

[21] B Subanya, R R Rajalaxmi, A novel feature selection algorithm for heart disease classification, International

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Journal of Computational Intelligence and Informatics, vol. 4: no. 2, July - September 2014.

[22] Eleyan D. Ant Colony Optimization based Feature Selection in Rough Set Theory[J]. Isaet Org, 2013(2). Chen, Y., Miao, D., & Wang, R. (2010). A rough set approach to feature selection based on ant colony optimization. Pattern Recognition Letters, 31(3), 226-233.

[23] Emary E, Zawbaa H M, Hassanien A E. Binary ant lion approaches for feature selection[J]. Neurocomputing, 2016, 213:54-65.

[24] Hu B, Dai Y, Su Y, et al. Feature selection for optimized high-dimensional biomedical data using the improved shuffled frog leaping algorithm[J]. IEEE/ACM transactions on computational biology and bioinformatics, 2016.

[25] Mafarja M M, Mirjalili S. Hybrid Whale Optimization Algorithm with simulated annealing for feature selection[J]. Neurocomputing, 2017, 260:302-312.

[26] Salcedo-Sanz S, Pastor-Sánchez A, Ser J D, et al. A Coral Reefs Optimization algorithm with Harmony Search operators for accurate wind speed prediction[J]. Renewable Energy, 2015, 75:93-101.

[27] Camacho-Gómez C, Wang X, Pereira E, et al. Active vibration control design using the Coral Reefs Optimization with Substrate Layer algorithm[J]. Engineering Structures, 2018, 157:14-26.

[28] Chu T H, Nguyen Q U, O’Neill M. Tournament Selection Based on Statistical Test in Genetic Programming[M]// Parallel Problem Solving from Nature – PPSN XIV. Springer International Publishing, 2016:303-312.

[29] Belew R K, McInerney J, Schraudolph N N. Evolving networks: Using the genetic algorithm with connectionist learning[C]//In. 1990.

[30] Ahmed M A, Alkhamis T M. Simulation-based optimization using simulated annealing with ranking and selection[J]. Computers & Operations Research, 2002, 29(4): 387-402.

[31] Zhu Z, Ong Y S, Dash M. Markov blanket-embedded genetic algorithm for gene selection[J]. Pattern Recognition, 2007, 40(11):3236-3248.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

- A novel framework based on an improved Coral Reefs Optimization is proposed, which can be applied to feature selection for biomedical data.
- Tournament selection strategy is employed to produce initial coral reef populations.
- Simulated Annealing (SA) is integrated with CRO to enhance the search performance of the original CRO algorithm.
- For various high dimensional biological datasets, we conduct experiments and 10-fold-CV is adopted to verify the effectiveness of proposed algorithm. At the same time we implemented all the comparison algorithms

ACCEPTED MANUSCRIPT