Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2015, Article ID 794586, 12 pages
http://dx.doi.org/10.1155/2015/794586

Research Article

Dimensionality Reduction in Complex Medical Data: Improved Self-Adaptive Niche Genetic Algorithm

Min Zhu,[1,2] Jing Xia,[1] Molei Yan,[3] Guolong Cai,[3] Jing Yan,[3] and Gangmin Ning[1]

[1]*Department of Biomedical Engineering, Zhejiang University, 38 Zheda Road, Hangzhou, Zhejiang 310027, China*
[2]*Guizhou Key Laboratory of Agricultural Bioengineering, Guizhou University, Guiyang, Guizhou 550025, China*
[3]*Zhejiang Hospital, Hangzhou, Zhejiang 310058, China*

Correspondence should be addressed to Gangmin Ning; gmning@zju.edu.cn

Received 24 July 2015; Revised 24 September 2015; Accepted 4 October 2015

Academic Editor: Anne Humeau-Heurtier

With the development of medical technology, more and more parameters are produced to describe the human physiological condition, forming high-dimensional clinical datasets. In clinical analysis, data are commonly utilized to establish mathematical models and carry out classification. High-dimensional clinical data will increase the complexity of classification, which is often utilized in the models, and thus reduce efficiency. The Niche Genetic Algorithm (NGA) is an excellent algorithm for dimensionality reduction. However, in the conventional NGA, the niche distance parameter is set in advance, which prevents it from adjusting to the environment. In this paper, an Improved Niche Genetic Algorithm (INGA) is introduced. It employs a self-adaptive niche-culling operation in the construction of the niche environment to improve the population diversity and prevent local optimal solutions. The INGA was verified in a stratification model for sepsis patients. The results show that, by applying INGA, the feature dimensionality of datasets was reduced from 77 to 10 and that the model achieved an accuracy of 92% in predicting 28-day death in sepsis patients, which is significantly higher than other methods.

1. Introduction

Clinical decision system is able to aid in diseases diagnosis and predict the clinical outcomes in response to treatment [1, 2]. For the diagnosis of sepsis, a number of scoring systems have been proposed, such as the Acute Physiology and Chronic Health Evaluation (APACHE), Sequential Organ Failure Assessment (SOFA), and Clinical Pulmonary Infection Score (CPIS) [1, 3]. They are challenged because traditional markers of infection mislead and there is lack of better evaluation methods for prognosis [1, 4–6]. To improve the outcome of treatments, diagnostic models are needed to accurately predict the development of sepsis as well as stratify its severity [7].

However, the clinical data of sepsis involved in diagnostic models are usually high dimensional. High-dimensional datasets increase the complexity of classification and reduce the effect of models [8]. Thus, before building models, it is necessary to reduce the data dimension while retaining essential information of the original data. Feature extraction and feature selection are the main methods in dimensionality reduction [2, 9].

(A) Feature Extraction. Feature extraction transforms the original feature space into a new one of lower dimension. Algorithms like Principal Component Analysis (PCA), Multidimensional Scaling (MDS), and Independent Component Analysis (ICA) are widely used for feature extraction. However, ICA and PCA are linear projection methods, and if the feature vectors distribute along a nonlinear manifold in a high-dimensional space, they might lead to classification errors [10, 11]. Besides, MDS is sensitive to undersampling datasets and has difficulty in dealing with defect data [12]. Furthermore, PCA, MDS, and ICA will generate new parameters after dimensionality reduction, and the significance of the new parameters is not always interpretable.

(B) Feature Selection. Feature selection is a kind of process that selects an optimal feature subset from the original features, which retains sufficient information [13]. Currently, quite

a lot of feature selection algorithms have been developed, such as Genetic Algorithms (GAs), Support Vector Machines (SVM) Wrapper, Sparse Generalized Partial Least Squares Selection (PLS), and Particle Swarm Optimization (PSO) [14–17]. Among them, GAs are popularly utilized. However, in some multimodal optimization problems, GAs failed to maintain multiple global or local optima [13]. Thus many efforts have been made to improve the ability of GAs in achieving multiple peak solutions, by adding scaling fitness and adjusting fitness competence rule [18].

(a) GAs. Genetic Algorithms have been used to reduce the numbers of features in datasets [19–21]. Genetic Algorithm Pipe Network Optimization Model (GENOME) has been applied to optimize the design of new looped irrigation water distribution networks [22]. An online web-based feature selection tool (DWFS) was developed according to the GA-based wrapper paradigm [23]. However, when using GAs [24–26], it is difficult to handle problems such as nonlinear, singular, and multimodal ones. The key issue is that the population is easily trapped in a limited number of solutions; and premature solutions have no capability to obtain better results [18]. Therefore, the Niche Genetic Algorithms (NGAs) are introduced to build a better environment to resolve the problem.

(b) NGAs. The capability to locate multiple loci often permits NGAs to be robust and effective in solving multimodal optimization problems [27–29]. The Twin-space Crowding Genetic Algorithm (TCGA) and Game-Theoretic Genetic Algorithm (GTGA) are introduced in the literature [18, 30]. The reported work [31] showed that the Nondominated Sorting Genetic Algorithm (NSGA) lacks elitism and needs to specify the sharing parameter [32]. However, most niche methods require prior knowledge such as the niche radius or the distance threshold. Accordingly, the niche distance is either set randomly or set as fixed value in advance. These technologies are unable to adaptively obtain the niche distance following evolution and prone to eliminate the potentially excellent individuals [33, 34].

To address the problems, we proposed Improved NGA (INGA) algorithm with embedded self-adaptive niche-culling mechanism for dimensionality reduction. Since MDS and PCA are the typical feature extraction algorithms while GA and NGA are the typical feature selection algorithms, we compared the dimension reduction results of them with INGA to verify the validity of INGA in dimension reduction. By applying INGA, the improvement in the accuracy rate of sepsis diseases classification is noteworthy, while the data dimension is reasonably reduced.

2. Method

The idea of NGA is applying the biological concept of a niche to evolutionary computations. It shows a survival environment with a prespecified distance parameter L. The L of NGA is set in advance, only allowing a single excellent individual in this distance. NGA has the following main disadvantages.

(1) A fixed distance parameter affects the convergence rate. If the value of L is too large, there will be lots of individuals within this distance and they need to be culled. This will lower the convergence rate. In contrast, if the value of L is too small, there are no sufficient individuals and this will lead to premature convergence.

(2) Single individual will inhibit potential individuals. Within the distance L, only one single excellent individual is allowed and it will cause the elimination of potentially excellent individuals and make the result of the dimension reduction too large.

(3) The diversity of the subpopulations is insufficient. Population diversity is closely related to subpopulations scale, but the subpopulations scale of NGA is set in advance and cannot be adjusted. It is difficult to find an optimum scale of subpopulations. As a result, if the subpopulations scale is too large, the diversity of the population is easy to be destroyed; on the contrary, the additional calculation of the algorithm will be increased.

To address these problems, we developed Niche Elimination Operation, as shown in the part (A). Afterwards, INGA is constructed, as shown in part (B) (Figure 2).

(A) Niche Elimination Operation

(a) Self-Adaptive Survival Distance. The distance parameter L is designed to be self-adaptive with the Euclidean distance among individuals of each generation to avoid the convergence problem caused by preset L:

$$D = \left\| X_i - X_j \right\| = \sqrt{\sum_{k=1}^{\text{len}} \left(x_{ik} - x_{jk} \right)^2}, \quad i, j \in \{1, 2, \ldots, M\}, \ i \neq j. \tag{1}$$

X_i and X_j are two individuals of the current population, which are made up of loci genetics. M is the number of individuals in the current population. len is the number of loci, which is used to form and evaluate the lengths of individuals. x_{ik} and x_{jk} are the values of loci. The distance parameter L is calculated by

$$L = \min \{D\}. \tag{2}$$

Because individuals of each generation are different and the values of the distance parameter vary with generation, a reasonable distance parameter will be obtained in the evolutionary process of each generation to get a better niche environment.

(b) Similarity Criterion. Allowing one single excellent individual within L, this will cause the elimination of potentially excellent individuals which may not be similar to the retained excellent. So, within the distance parameter L, the similarities of biallelic loci are used to judge the similarity of

the individuals and determine whether the individuals should be retained.

The similarity of biallelic loci and average similarity between two individuals are given by the following two equations:

$$\mathrm{SD}\left(X_i, X_j\right) = \sum_{k=1}^{\mathrm{len}} \frac{\mathrm{num}\left(X_{ik} == X_{jk}\right)}{\mathrm{len}}, \quad i \in \{1, 2, \ldots, M\}, \ j \in \{i+1, i+2, \ldots, M\}, \tag{3}$$

where $\mathrm{SD}(X_i, X_j)$ represents the similarity between two individuals, X_i and X_j. $\mathrm{num}(X_{ik} == X_{jk})$ is the number of the same allele value of two individuals. Consider

$$\mathrm{MSD}_i = \frac{\sum_{j=i+1}^{M} \mathrm{SD}\left(X_i, X_j\right)}{\mathrm{len} * (M-1)}, \quad i \in \{1, 2, \ldots, M\}, \ j \in \{i+1, i+2, \ldots, M\}. \tag{4}$$

MSD_i represents the average similarity between the ith individual and the others. When $\|X_i - X_j\| < L$, the similarity between two individuals will be distinguished. If the similarity is larger than the average similarity, the individual that has a lower fitness will be given a penalty function, as shown in the following equation. Otherwise, the lower fitness individuals can be retained:

$$f_j'(X) = f_j(X) * P, \tag{5}$$

where $f_j(X)$ is the original fitness of the individual, $f_j'(X)$ is the new fitness, and P is the penalty function (usually 10^{-30}). This method can reduce the elimination of individuals.

(c) Maintain Population Diversity. To maintain the diversity of the population, the scale of the subpopulations should be controlled. So (6) and (7) are designed with a memory pool of optimal individuals to limit the scale for the subpopulations of each generation:

$$f(t) = \sum_{i=1}^{M_{(t)}} \frac{f_i(t)}{M_{(t)}}, \tag{6}$$

where $f(t)$ represents the average fitness value of generation t, $f_i(t)$ represents the fitness of individual i in generation t, and $M_{(t)}$ is the scale of the population in generation t. Thus, the scale of subpopulations in generation $t+1$ is $M_{(t+1)}$. This is calculated as

$$M_{(t+1)} = M_{(t)} \cdot f(t) \cdot \frac{t}{\sum_{i=1}^{t} f(i)}. \tag{7}$$

A memory pool of optimal individuals is designed to exchange excellent evolutionary individuals. The operation increases the possibility of obtaining more excellent individuals, and to some extent, avoids the problem of premature convergence during the evolutionary process of a single population. The individuals of general $t+1$ are sorted by fitness, and the formers N are put into the memory pool.

Through the result of $M_{(t+1)}$, the ability of maintaining the population diversity, $d(p)$, is designed as in the following two equations. The smaller the value of $d(P)$ is, the higher its population diversity is:

$$d(p) = \frac{\sum_1^t d(P)_t}{t}, \tag{8}$$

where $d(P)_t$ is the capability to maintain the population diversity in generation t. And $d(P)_t$ is designed as follows:

$$d(P)_t = \frac{1}{l \cdot M_{(t)}} \sum_{j=1}^{l} \max\left\{\sum_{i=1}^{M_{(t)}} \left(1 - a_{ij}\right), \ \sum_{i=1}^{n} a_{ij}\right\}, \tag{9}$$

where l is the length of the individual encoding, $M_{(t)}$ is the scale of the population in generation t, and a_{ij} is the jth loci of the ith individual.

(B) Flowchart of INGA

Step 1 (calculate fitness). At first, M initial individuals are produced at random. Usually, it takes the reciprocal of the sum of error square of the classifier test set data as fitness function [33] in order to fully reflect the advantage of controlling errors by combining INGA with classifier:

$$f(X) = \frac{1}{\sum_{i=1}^{n} \left(\widehat{t}_i - t_i\right)^2}, \tag{10}$$

where $\widehat{t}$ is the predicted value of test set, t is the true value of test set, and n is the sample number of test set. Individuals are sorted by fitness in descending order, and the former N individuals are remembered in the memory pool ($N < M$).

Step 2 (Niche Elimination Operation to produce excellent initial individuals). In this step, the excellent initial individuals $M_{(t)}$ are produced, as shown in Figure 1.

(a) *Self-Adaptive Survival*. First, calculate the Euclidean distance D between X_i and X_j according to (1). Second, calculate self-adaptive survival distance L according to (2).

(b) *Similarity Criterion*. Judge the similarity of the individuals within the distance L according to the method of allele contrast, so as to determine whether the individual should be retained. When $\|X_i - X_j\| < L$, the similarity of biallelic loci and average similarity between two individuals are compared. If they are not similar, the individual of lower fitness needs not to be eliminated. The similarity of biallelic loci $\mathrm{SD}(X_i, X_j)$ and average similarity MSD_i between two individuals are given by (3) and (4). When $\mathrm{SD}(X_i, X_j) > \mathrm{MSD}_i$, then $f_j(X)$ is punished, using a penalty function $f_j'(X) = f_j(X) * P$ according to (5). If not, the individual with lower fitness will be retained. On the other hand, when $\|X_i - X_j\| > L$, the individual with lower fitness will be retained.

(c) *Maintaining Population Diversity*. According to (7), the number of subpopulations $M_{(t+1)}$ is calculated.

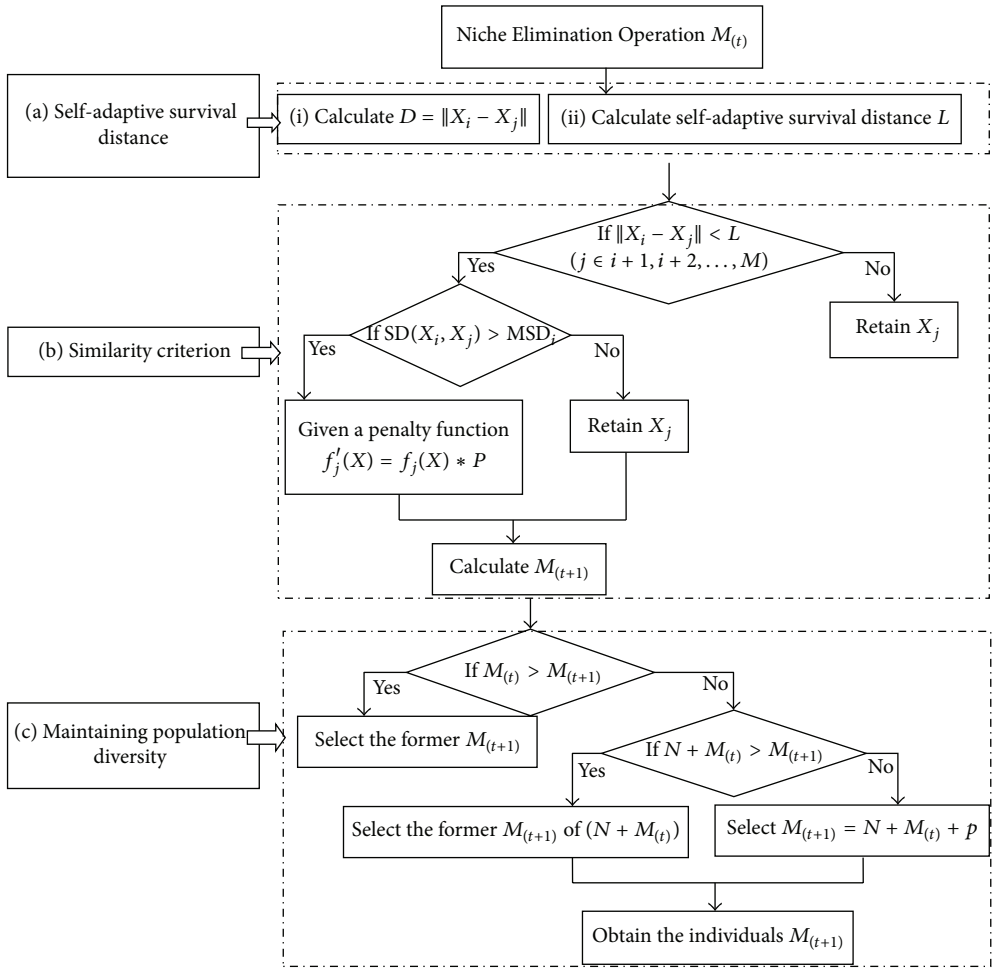


Figure 1: Niche Elimination Operation.

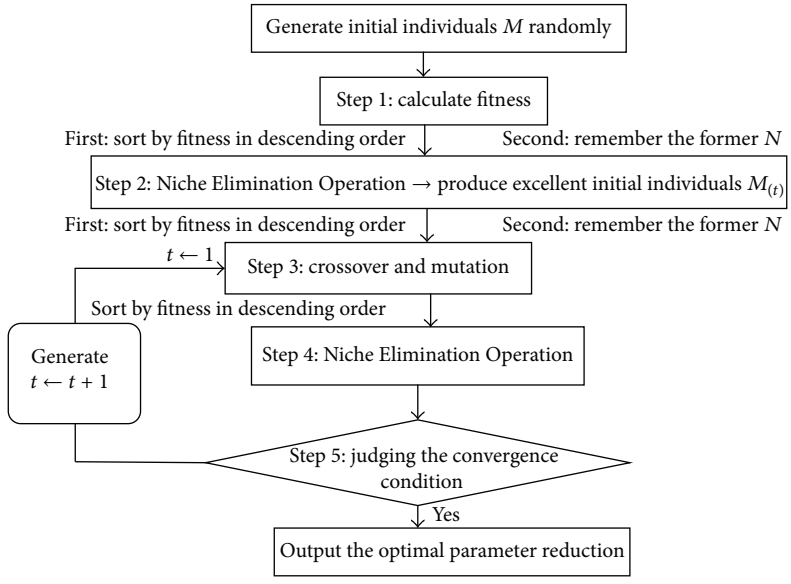


Figure 2: Flowchart of INGA.

Individuals are sorted by fitness in descending order, if the scale of the existing subpopulation $M(t)$ is larger than $M(t+1)$, select the individuals $M_{(t+1)}$; otherwise, N individuals are merged in the memory pool with the existing subpopulations and sorted by fitness in descending order; when $N + M_{(t)} > M_{(t+1)}$, the former individuals $M_{(t+1)}$ of $(N + M_{(t)})$ are selected; when $N + M_{(t)} < M_{(t+1)}$, P individuals will be generated randomly; individuals $M_{(t+1)}$ are selected, on the condition that $M_{(t+1)} = N + M_{(t)} + p$. Through this method, the initial population will have a higher average fitness and will be conducive to the evolution of population towards the solution of the problem.

Step 3 (self-adaptive crossover and mutation operation). Considering the probability of crossover and mutation, it is too small to escape from making the system fall into the local optimal solution, and if it is too large, it can escape from the local optimal solution but is prone to instability and convergence because the count of crossover and mutation is so frequent. In order to improve this shortcoming, the equations of self-adaptive crossover (P_c) and mutation probability (P_m) are used [35, 36]:

$$P_c = \begin{cases} P_{c1} - \dfrac{P_{c1} - P_{c2}}{f_{\max} - f_{\text{avg}}} \left(f' - f_{\text{avg}}\right) & f' \geq f_{\text{avg}} \\ P_{c1} & f' < f_{\text{avg}}, \end{cases} \tag{11}$$

$$P_m = \begin{cases} P_{m1} - \dfrac{P_{m1} - P_{m2}}{f_{\max} - f_{\text{avg}}} \left(f - f_{\text{avg}}\right) & f \geq f_{\text{avg}} \\ P_{m1} & f < f_{\text{avg}}. \end{cases} \tag{12}$$

$f_{\max}$ is the maximum fitness value; f_{avg} is the average fitness value of each population; f' is the larger fitness value of the two individuals crossing; and f is the fitness value of individuals of mutation. P_{c1}, P_{c2} are, respectively, the crossover probability value of two individuals; P_{m1} and P_{m2} are the mutation probability values of two individuals.

Step 4 (Niche Elimination Operation). After the self-adaptive crossover and mutation operation, put the new individual into the Niche Elimination Operation again to obtain the optimal individual, as shown in Figure 1.

Step 5 (judging the termination condition). If it does not meet the termination condition, then update the counter t as $t + 1$ and make the population in Step 4 be the new next generation population, and then go to Step 2. If the termination condition is satisfied, output the optimal dimensionality reduction parameters selected.

3. Dataset Description

Experiments are conducted on a sepsis dataset, for which data are gathered from Zhejiang Hospital. The goal of the classifier was to determine, based upon the test results provided, whether a patient should be diagnosed as 28-day death [37]. The number of samples in the two classes was balanced. The training set contained 124 negative (28-day death) cases and 173 positive cases. Likewise, the testing set consisted of 77 negative samples and 123 positive ones. Data are organized in a table with 77 columns for attributes of patients and 497 rows for specific samples. There are missing values in this table because some questions have not been answered, so we replaced them with 0. There is not any correlation among attributes, and this creates an orthogonal space for using Euclidean distance. All samples include the same number of attributes [13].

4. Experimental Setup

This work used the PCA, MDS, NGA, and INGA to reduce the dimensionality of the dataset, and the selected algorithms were also combined with three classic classifiers, Random Forest (RF), Support Vector Machine (SVM), and Back Propagation (BP). The experimental setup is as follows.

Set the Initial Population Scale. The literature [37, 38] suggests that an optimal initial population should number from 20 to 100; the present work takes 90 as the initial population M, considering the computation time and the range of the search. The stored individuals N in Niche Genetic Algorithm are usually selected as one-thirds of population scale. The probability of crossover is determined by (11), and the mutation probability is determined by (12).

Set the Encoding. The data are organized in a table with 77 columns for attributes of patients and each bit is assigned to one feature; thus the encoding length is designed as 77. If the ith bit equals 1, then the ith feature is involved in classification; otherwise, the corresponding feature is not involved, as shown in Figure 3.

Set the Convergence Condition. The evolutional generation is set to 100 according to the previously published works [13, 37]. The fitness function is the reciprocal of the sum of the prediction error square of the model. Convergence is achieved when the largest and least fitness values are equivalent. This paper adopts the maximum evolutional generation and convergence degree of the population to construct the condition of algorithm convergence: end the calculation when it can meet one of the two conditions; namely, the evolutional generation reaches the preset values or population convergence appears [36].

Set the Experiment Running Time. The experiment used k-fold cross-validation, 80% of the samples were randomly selected as the training set, and the rest were used as the test set. The experiment was repeated 100 times [39].

5. Result

The clinical manifestation of the sepsis disease is complicated, and it is difficult to accurately determine the 28-day mortality. This study applies the improved self-adaptive Niche Genetic Algorithm to the diagnosis process of septic 28-day mortality, using dimensionality reduction to obtain the optimal feature parameters and improve the diagnostic precision.

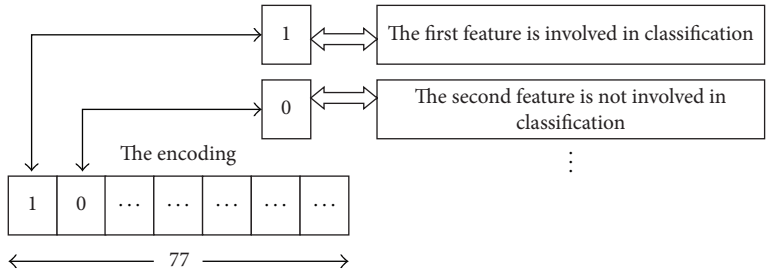


Figure 3: The relationship between encoding and features.

Table 1: $d(P)$: the ability to maintain population diversity. $d(P)$ was run for 20, 50, and 100 generations, respectively, in GA, NGA, and INGA.

Generation	GA	NGA	INGA
20	0.5635	0.5213	0.4812
50	0.6271	0.5748	0.5248
100	0.6963	0.6147	0.5629

Here, premature state, population distribution, accuracy of classification, and robustness have been used to measure the quality of the algorithms.

(A) Premature State. Avoiding premature state is a standard of the algorithms; premature means that the performance is as follows: (a) the population diversity is reduced, (b) the convergence ability is low, and (c) the convergence rate is low. Thus, we used these factors to measure whether the algorithms were premature or not.

(a) Population Diversity. The Schaffer function, presented as (13), is used to generate data, and the results of population diversity, with $d(P)$ calculated with (9), are shown in Table 1:

$$f(x_1, x_2) = 0.5 - \frac{\sin^2\sqrt{x_1^2 + x_2^2} - 0.5}{\left(1 + 0.001\left(x_1^2 + x_2^2\right)\right)^2}, \quad -100 \le x_i \le 100 \ \ (i = 1, 2). \tag{13}$$

We can see from Table 1 that the value of $d(P)$ of INGA is smaller than that of GA and NGA under the condition of the same evolution generations, demonstrating the advantages of INGA in maintaining the population diversity.

(b) Convergence Ability. Convergence ability means the ability to obtain global optimal values when algorithm stops. We know from the properties of the Schaffer function that the global maximum is 1 and that two local maxima near the maximum value are 0.99028 and 0.96278. If the maximum value was larger than 0.999, we can judge the convergence appearance, and the global solution is obtained. When local maxima values are obtained, we can judge that there is no convergence, as only the local solution is obtained. Thus, GA, NGA, and INGA are used to obtain the maximum value of the Schaffer function, as shown in Table 2.

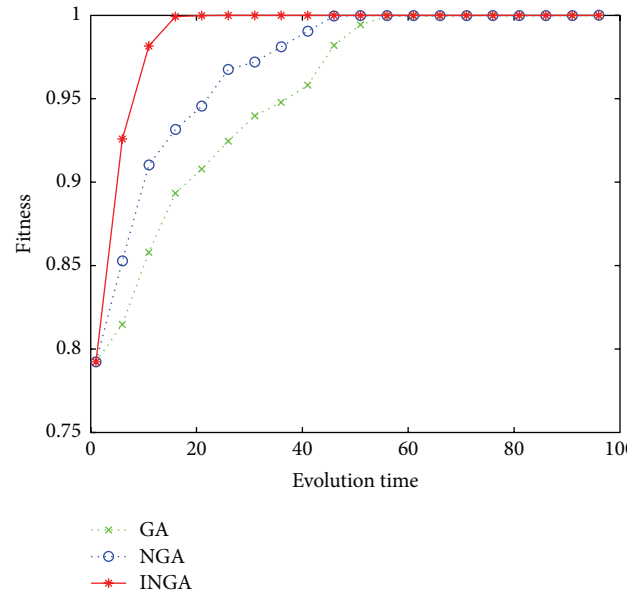


Figure 4: Convergence curves.

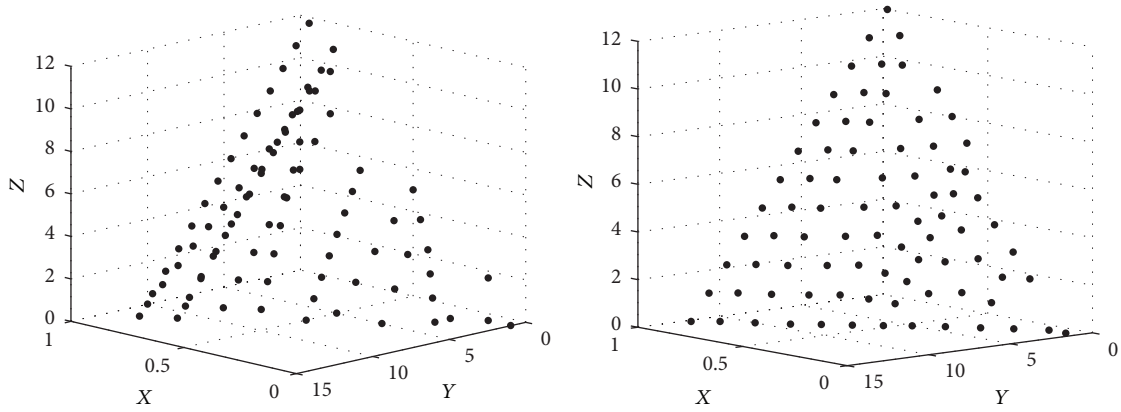
From the data in Table 2, we can see that, in the 10 independent experiments, it is easier for GA and NGA to fall into two local maxima. There are 10 times for INGA to search the global optimal value, there are 7 times for NGA to search the global optimal solution, and GA only has 4 times, which means that there is a certain gap between the ability of these two algorithms to search for the global optimal solution compared with INGA.

(c) Convergence Rate. The comparison of convergence curves among GA, NGA, and INGA is shown in Figure 4. We can see from Figure 4 that INGA has the fastest convergence rate. It has converged to the average fitness by the 20th generation. The remaining two algorithms converged to the average fitness by the 42th and 67th generations, respectively.

(B) Population Distribution. In Section 2, self-adaptive survival distance is used to set up the distance of NGA, and criterion similarity is used to determine whether the individual is retained or not. Both of them constitute the population distribution. So the figure of population distribution is built to assess the effect of the self-adaptive survival distance and criterion similarity methods.

Table 2: Convergence of the Schaffer function.

Execution count	GA		NGA		INGA	
	Optimal value	Whether converges	Optimal value	Whether converges	Optimal value	Whether converges
1	0.9903	N	0.9903	N	0.9995	Y
2	0.9632	N	0.9998	Y	1	Y
3	1	Y	1	Y	0.9998	Y
4	0.9619	N	0.9625	N	1	Y
5	0.9991	Y	0.9991	Y	0.9995	Y
6	0.9631	N	0.9995	Y	1	Y
7	0.9631	N	1	Y	0.9998	Y
8	0.9992	Y	0.9628	N	1	Y
9	0.9617	N	0.9982	Y	1	Y
10	0.9996	Y	1	Y	0.9998	Y



(a) NGA (b) INGA

Figure 5: Individuals' distribution, where the x-axis represents the fitness of the individuals and the y- and z-axes represent the Euclidean distance between the individuals.

Table 3: The number of feature parameters after dimensionality reduction.

	PCA	MDS	NGA	INGA
BP	17 ± 2	27 ± 3	21 ± 3	15 ± 2
SVM	2 ± 10	26 ± 4	20 ± 4	16 ± 3
RF	21 ± 4	22 ± 3	23 ± 3	10 ± 2

Figure 5 is the population distribution within the niche distance. It shows that the final population obtained by INGA can be more uniformly distributed; thus self-adaptive survival distance and similarity criterion designed in this paper is adaptive.

(C) Dimensionality Reduction and Classification. To assess the performance of dimensionality reduction, PCA, MDS, NGA, and INGA were embedded in the BP, SVM, and RF classifiers to carry out the classification diagnosis. The accuracy of classification and ROC curve diagram are shown as follows.

(a) Accuracy of Classification. The number of feature subsets before and after dimensionality reduction is shown in Table 3. It is shown that INGA has better control over the number of feature subsets than other dimensionality reduction methods, as a smaller number of feature subsets were obtained by INGA. However, considering the number of feature subsets alone is not enough, as the classification accuracy should be combined. The classification accuracies before and after dimensionality reduction are shown in Figure 6. It is noticed that the accuracy increased obviously after the dimensionality reduction; the highest accuracy was obtained by RF-INGA.

(b) The ROC Curves. The receiver operating characteristic (ROC) curve and area under the curve (AUC) are shown in Figures 7 and 8.

From Figures 7 and 8, we can see that INGA yields a better result and that the covered areas of ROC offer an obvious improvement compared with PCA, MDS, and NGA. At the same time, the highest AUC was obtained by the RF classifier after INGA dimensionality reduction.

(D) Robustness. The robustness of the algorithm was tested by introducing random noise in the data. The k-fold cross-validation method was used to compare the effects of noise. 5% of the samples, selected randomly from the training

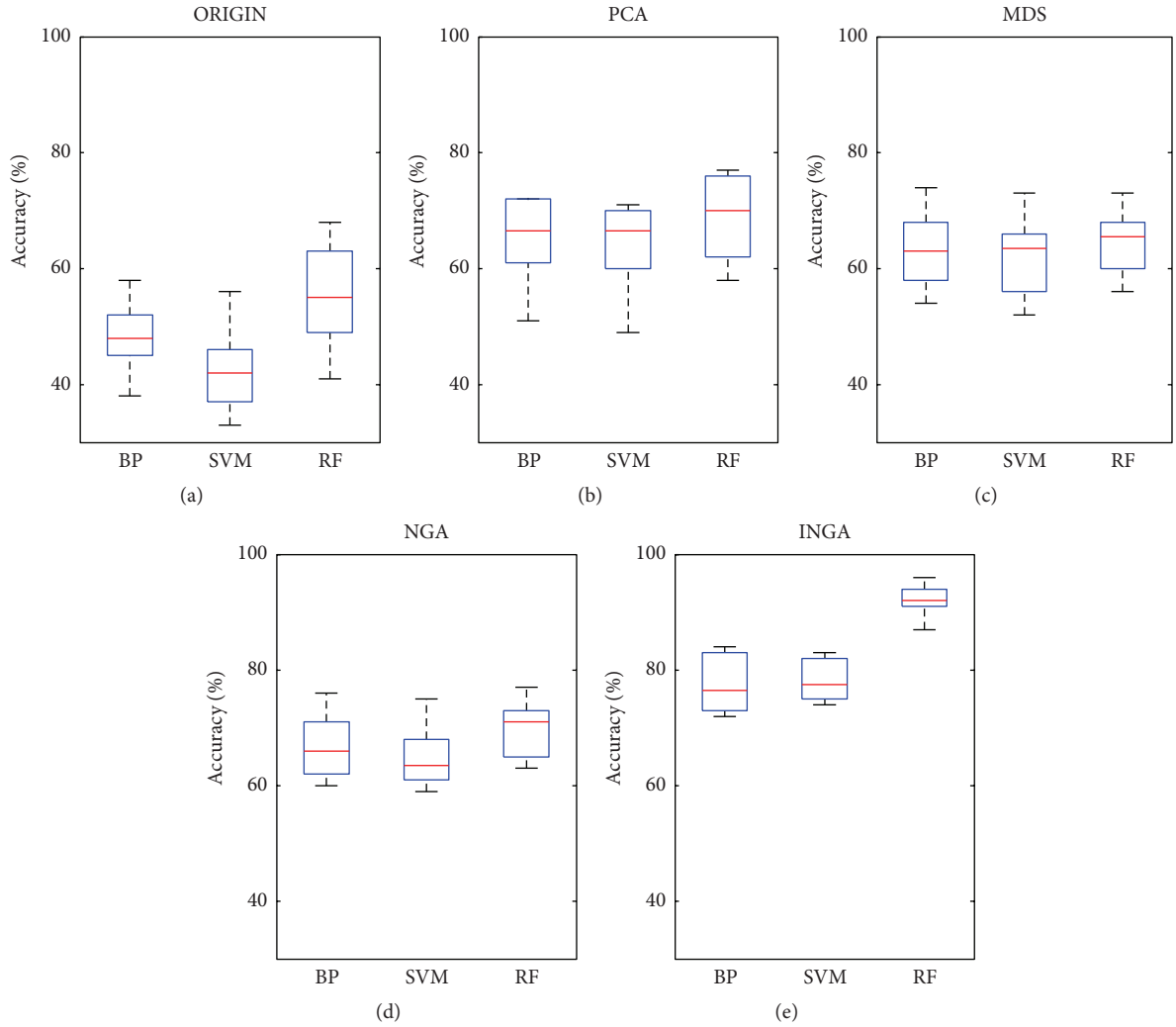


Figure 6: Classification accuracy (%). (a) is the result before dimensionality reduction and (b)–(e) are the result after dimensionality reduction.

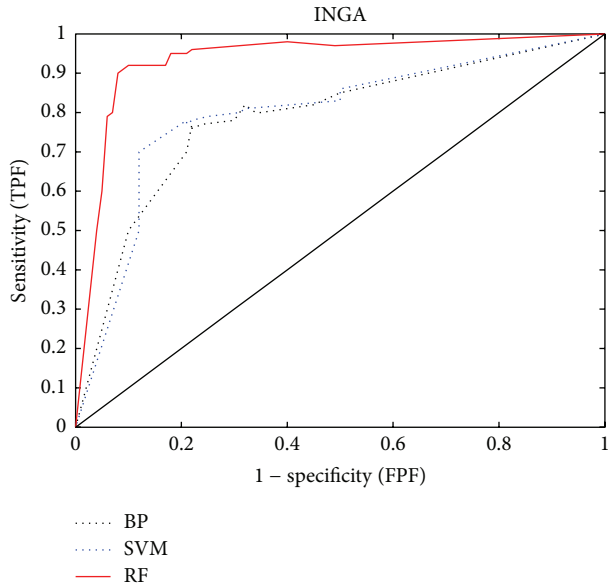


Figure 7: ROC curves. Classification using BP, SVM, and RF based on the INGA dimension reduction algorithm.

set, and their labels are changed, used as noise samples. The operation was repeated 100 times, and the average value was taken to compare the classification accuracy.

From Figure 9, we can see that noise poses a significant effect on the dimensionality reduction methods of PCA and MDS. In comparison with Figure 6, the accuracy of the three classifiers decreased by 18% to 35%; on the contrary, INGA is less affected by the noisy conditions, and the accuracy of the three classifiers with INGA only decreased by 3% to 13%. The robustness of the INGA algorithm is strengthened, and its antinoise ability is the best, especially when it is combined with RF.

6. Discussion

The integrated feature selection algorithms and classification accuracy were valid on clinical sepsis data. INGA exhibited advantages in feature selection over other approaches, and, moreover, INGA-RF obtained classification accuracy higher than 90% in identifying the death of sepsis patients, showing the best performance of all of the techniques and using only 10 features.

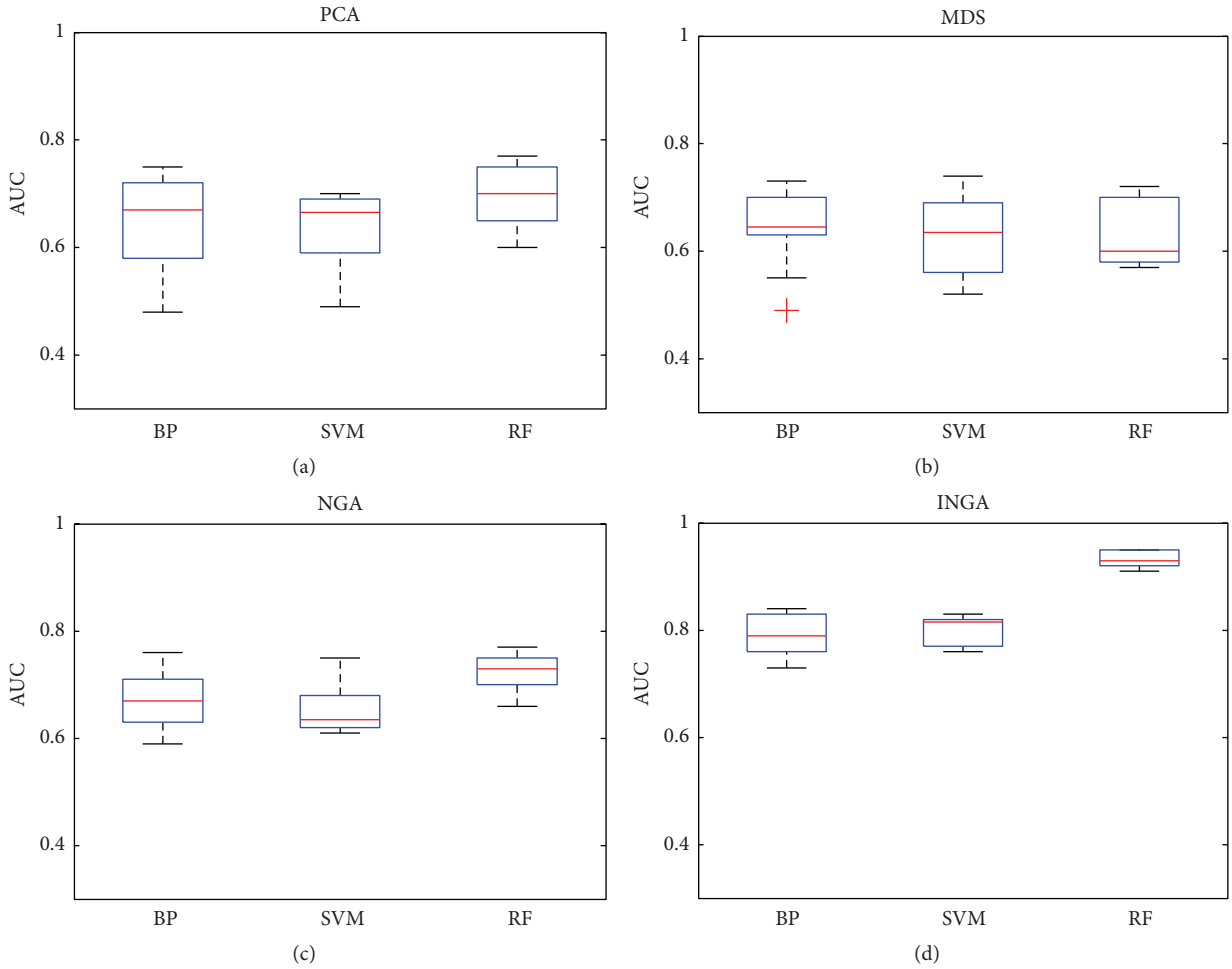


Figure 8: Area under the curve (AUC) of the algorithm.

The present work has proposed an improved INGA algorithm to resolve the premature state in traditional GA and NGA, which are characterized as having reduced population diversity, weak convergence ability, and low convergence rate. As shown in Table 1, regarding $d(P)$, a measure of population diversity, INGA has the smallest value of 0.5629 as compared with GA and NGA. As shown in Table 2, for 10 independent experiments, INGA achieved the global optimum in all experiments, while NGA and GA succeeded in only 7 and 4 experiments, respectively. Figure 4 shows the convergence rate estimated by the generations of convergence. INGA had the fastest convergence rate with 20 generations, while 42 and 67 generations were required for GA and NGA, respectively. These findings suggest that INGA is superior overall to the other methods.

The dominating performance of INGA in avoiding premature convergence is owed to the following improvements in the work: (i) the introduction of the self-adaptive survival distance: differing from the conventional methods, the survival distance is automatically adjusted in the evolutionary process of each generation; this ensures reasonable distance parameters and leads to an adaptive niche environment; this approach can obtain more reasonable individuals with excellent global optimization ability and high convergence speed; (ii) the application of a similarity criterion that retains more reasonable individuals: the similarity of biallelic loci was used to decide whether the individuals in the neighborhood should be retained; this approach can harvest more reasonable individuals, increasing the possibility of finding the global optimal solution; and (iii) the use of a memory pool for optimal individuals: a pool was designed to reserve and exchange excellent evolutionary individuals for each generation; this maintains the diversity of the population and increases the quantity of excellent individuals; to some extent, it also avoids the problem of premature convergence during the evolutionary process of a single population.

The testing results on clinical sepsis cases show that, combined with INGA, three types of classifiers achieved the accuracies in predicting 28-day death of 92% (RF), 78% (SVM), and 77% (BP), respectively. In contrast, the highest accuracy of the classifiers employing NGA, PCA, and MDS is only 70%. This suggests that INGA is effective in improving the performance of classifiers for complex clinical datasets.

However, it is worth pointing out that the present work has some limitations. First, the validity of INGA was only tested in sepsis patients. Although the algorithm is generally functional, it is necessary to investigate the effectiveness of INGA on further datasets. Second, the coherence between

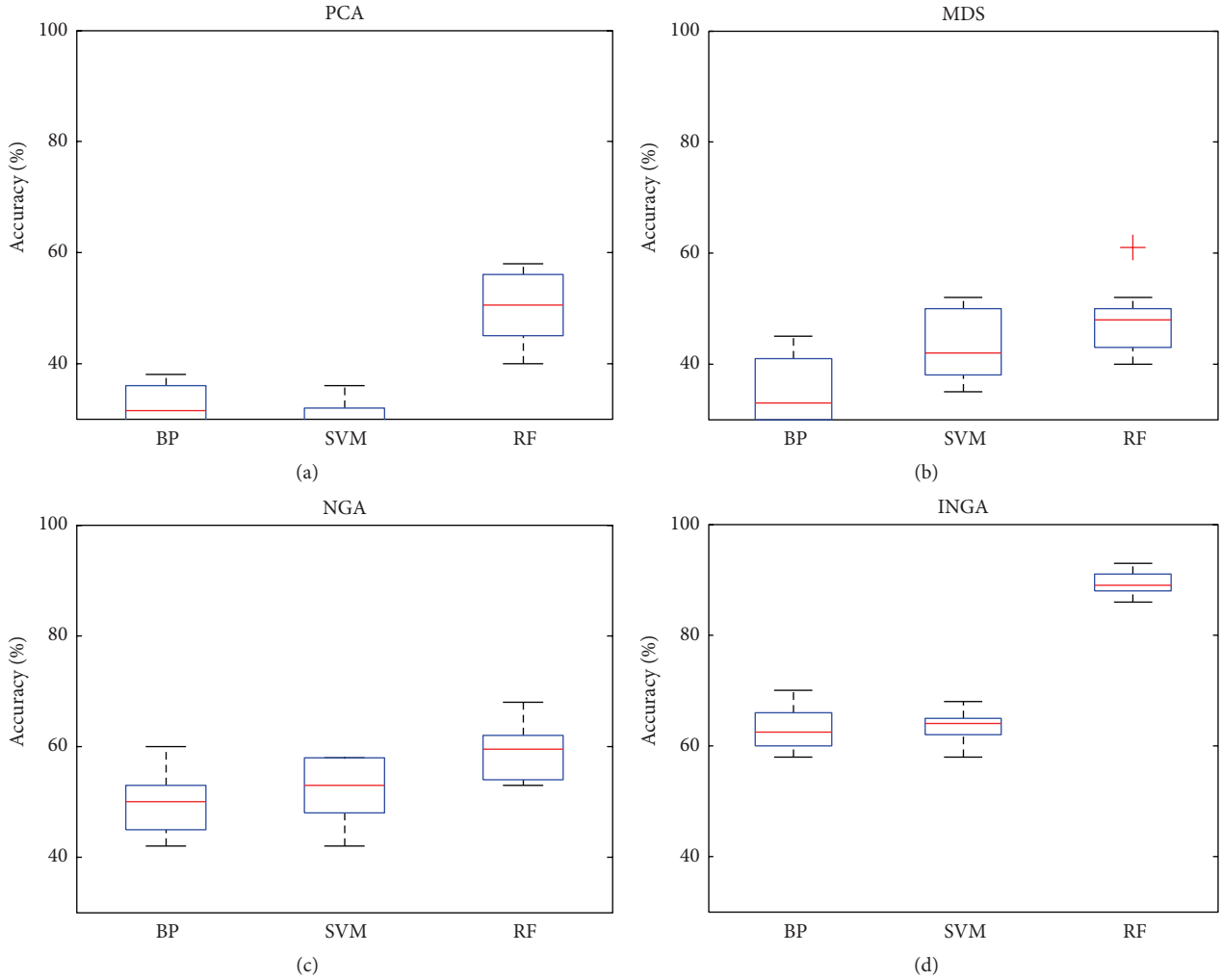


Figure 9: Robustness (%) of the algorithm.

INGA and the classifiers remains unclear. Our work revealed that the RF method with embedded INGA is mostly satisfied. One question that may arise is how to figure out the optimum combination of the dimension reduction algorithm and classifier. This question is out of the scope of the current work, which is focused on the dimension reduction. However, it should be clarified in a further study.

7. Conclusion

This paper proposed an improved algorithm for feature reduction in high-dimensional data. The methods were imbedded in classifiers to predict the prognosis of sepsis patients based on complex clinical datasets. The results indicate that the improved NGA, INGA, is most effective in reducing the number of attributes and enhancing the convergence speed compared to other commonly used algorithms, such as PCA, MDS, and NGA. Moreover, INGA associated with RF to achieve the highest accuracy in assessing the severity of sepsis. This suggests that INGA has the potential for complex data processing, particularly for medical pattern recognition.

Conflict of Interests

Min Zhu, Jing Xia, Molei Yan, Guolong Cai, Jing Yan, and Gangmin Ning declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported by the National Nature Science Foundation of China (Grant 81271662), the Department of Science and Technology of Zhejiang Province (Grant 2011R50018), and the Ministry of Health of China (Grant 201202011).

References

[1] K. M. Ho, "Combining sequential organ failure assessment (SOFA) score with acute physiology and chronic health evaluation (APACHE) score to predict hospital mortality of critically ill patients," *Anaesthesia and Intensive Care*, vol. 35, pp. 515–521, 2007.

[2] T. Jirapech-Umpai and S. Aitken, "Feature selection and classification for microarray data analysis: evolutionary methods for

identifying predictive genes," *BMC Bioinformatics*, vol. 6, article 148, 2005.

[3] C. BalcI, H. Sungurtekin, E. Gürses, U. Sungurtekin, and B. Kaptanoglu, "Usefulness of procalcitonin for diagnosis of sepsis in the intensive care unit," *Critical Care*, vol. 7, no. 1, pp. 85–90, 2003.

[4] E. Silva, M. D. A. Pedro, A. C. B. Sogayar et al., "Brazilian sepsis epidemiological study (BASES study)," *Critical Care*, vol. 8, no. 4, pp. R251–R260, 2004.

[5] C.-C. Jenq, M.-H. Tsai, Y.-C. Tian et al., "RIFLE classification can predict short-term prognosis in critically ill cirrhotic patients," *Intensive Care Medicine*, vol. 33, no. 11, pp. 1921–1930, 2007.

[6] R. Seligman, M. Meisner, T. C. Lisboa et al., "Decreases in procalcitonin and C-reactive protein are strong predictors of survival in ventilator-associated pneumonia," *Critical Care*, vol. 10, no. 5, article R125, 2006.

[7] N. G. Morgenthaler, J. Struck, M. Christ-Crain, A. Bergmann, and B. Müller, "Pro-atrial natriuretic peptide is a prognostic marker in sepsis, similar to the APACHE II score: an observational study," *Critical Care*, vol. 9, no. 1, pp. R37–45, 2005.

[8] K. K. Bharti and P. K. Singh, "A three-stage unsupervised dimension reduction method for text clustering," *Journal of Computational Science*, vol. 5, no. 2, pp. 156–169, 2014.

[9] Z. Liu, T. Chai, W. Yu, and J. Tang, "Multi-frequency signal modeling using empirical mode decomposition and PCA with application to mill load estimation," *Neurocomputing*, vol. 169, pp. 392–402, 2015.

[10] J. Salmon, Z. Harmany, C.-A. Deledalle, and R. Willett, "Poisson noise reduction with non-local PCA," *Journal of Mathematical Imaging and Vision*, vol. 48, no. 2, pp. 279–294, 2014.

[11] J. Yin, Y. Wang, and J. Hu, "A new dimensionality reduction algorithm for hyperspectral image using evolutionary strategy," *IEEE Transactions on Industrial Informatics*, vol. 8, no. 4, pp. 935–943, 2012.

[12] M. Mignotte, "MDS-based multiresolution nonlinear dimensionality reduction model for color image segmentation," *IEEE Transactions on Neural Networks*, vol. 22, no. 3, pp. 447–460, 2011.

[13] M. L. Raymer, W. F. Punch, E. D. Goodman, L. A. Kuhn, and A. K. Jain, "Dimensionality reduction using genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 4, no. 2, pp. 164–171, 2000.

[14] K. Tanaka, T. Kurita, and T. Kawabe, "Selection of import vectors via binary particle swarm optimization and cross-validation for kernel logistic regression," in *Proceedings of the International Joint Conference on Neural Networks (IJCNN '07)*, pp. 1037–1042, Orlando, Fla, USA, August 2007.

[15] A. Ratnaweera, S. K. Halgamuge, and H. C. Watson, "Self-organizing hierarchical particle swarm optimizer with time-varying acceleration coefficients," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 3, pp. 240–255, 2004.

[16] S. A. Khan, M. Nazir, N. Riaz, and M. Khan, "Optimized features selection using hybrid PSO-GA for multi-view gender classification," *The International Arab Journal of Information Technology*, vol. 12, no. 2, pp. 183–189, 2015.

[17] R. Bello, Y. Gomez, A. Nowe, and M. M. Garcia, "Two-step particle swarm optimization to solve the feature selection problem," in *Proceedings of the 7th International Conference on Intelligent Systems Design and Applications (ISDA '07)*, pp. 691–695, Rio de Janeiro, Brazil, October 2007.

[18] C. Chen, T. Liu, and J. Chou, "A novel crowding genetic algorithm and its applications to manufacturing robots," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 3, pp. 1705–1716, 2014.

[19] F. E. Fassnacht, H. Latifi, and B. Koch, "An angular vegetation index for imaging spectroscopy data—preliminary results on forest damage detection in the Bavarian National Park, Germany," *International Journal of Applied Earth Observation and Geoinformation*, vol. 19, no. 1, pp. 308–321, 2012.

[20] G. C. Dandy, A. R. Simpson, and L. J. Murphy, "An improved genetic algorithm for pipe network optimization," *Water Resources Research*, vol. 32, no. 2, pp. 449–458, 1996.

[21] H. Latifi, A. Nothdurft, and B. Koch, "Non-parametric prediction and mapping of standing timber volume and biomass in a temperate forest: application of multiple optical/LiDAR-derived predictors," *Forestry*, vol. 83, no. 4, pp. 395–407, 2010.

[22] J. Reca and J. Martínez, "Genetic algorithms for the design of looped irrigation water distribution networks," *Water Resources Research*, vol. 42, no. 5, Article ID W05416, 2006.

[23] O. Soufan, D. Kleftogiannis, P. Kalnis, V. B. Bajic, and D. Gupta, "DWFS: a wrapper feature selection tool based on a parallel genetic algorithm," *PLoS ONE*, vol. 10, no. 2, Article ID e0117988, 2015.

[24] C.-T. Cheng, K. Fallahi, H. Leung, and C. K. Tse, "A genetic algorithm-inspired UUV path planner based on dynamic programming," *IEEE Transactions on Systems, Man, and Cybernetics Part C: Applications and Reviews*, vol. 42, no. 6, pp. 1128–1134, 2012.

[25] C.-C. Hsu, Y.-J. Chen, M.-C. Lu, and S.-A. Li, "Optimal path planning incorporating global and local search for mobile robots," in *Proceedings of the 1st IEEE Global Conference on Consumer Electronics (GCCE '12)*, pp. 668–671, IEEE, Tokyo, Japan, October 2012.

[26] C. Hocaoglu and A. C. Sanderson, "Planning multiple paths with evolutionary speciation," *IEEE Transactions on Evolutionary Computation*, vol. 5, no. 3, pp. 169–191, 2001.

[27] L. Qing, W. Gang, Y. Zaiyue, and W. Qiuping, "Crowding clustering genetic algorithm for multimodal function optimization," *Applied Soft Computing*, vol. 8, no. 1, pp. 88–95, 2008.

[28] A. A. Alugongo, "Multimodal problems, premature convergence versus computation effort in dynamic design optimization," in *Proceedings of the World Congress on Engineering*, vol. 3, London, UK, July 2011.

[29] B. Sareni, L. Krähenbühl, and A. Nicolas, "Niching genetic algorithms for optimization in electromagnetics. I. Fundamentals," *IEEE Transactions on Magnetics*, vol. 34, no. 5, pp. 2984–2987, 1998.

[30] A. Konak, S. Kulturel-Konak, and L. V. Snyder, "A game-theoretic genetic algorithm for the reliable server assignment problem under attacks," *Computers & Industrial Engineering*, vol. 85, pp. 73–85, 2015.

[31] K. Deb, S. Agrawal, A. Pratap, and T. Meyarivan, "A fast elitist non-dominated sorting genetic algorithm for multi-objective optimization: NSGA-II," in *Parallel Problem Solving from Nature PPSN VI*, vol. 1917 of *Lecture Notes in Computer Science*, pp. 849–858, Springer, Berlin, Germany, 2000.

[32] J. Horn, N. Nafpliotis, and D. E. Goldberg, "A niched Pareto genetic algorithm for multiobjective optimization," in *Proceedings of the 1st IEEE Conference on Evolutionary Computation*, vol. 1, pp. 82–87, IEEE, Orlando, Fla, USA, June 1994.

[33] A. E. I. Brownlee, O. Regnier-Coudert, J. A. McCall, S. Massie, and S. Stulajter, "An application of a GA with Markov network

surrogate to feature selection," *International Journal of Systems Science*, vol. 44, no. 11, pp. 2039–2056, 2013.

[34] B. Sareni and L. Krähenbühl, "Fitness sharing and niching methods revisited," *IEEE Transactions on Evolutionary Computation*, vol. 2, no. 3, pp. 97–106, 1998.

[35] C. W. Ho, K. H. Lee, and K. S. Leung, "A genetic algorithm based on mutation and crossover with adaptive probabilities," in *Proceedings of the Congress on Evolutionary Computation (CEC '99)*, vol. 1, IEEE, Washington, DC, USA, July 1999.

[36] T. Ingu and H. Takagi, "Accelerating a GA convergence by fitting a single-peak function," in *Proceedings of the IEEE International Fuzzy Systems Conference*, vol. 3, pp. 1415–1420, Seoul, Republic of Korea, August 1999.

[37] B. L. Miller and M. J. Shaw, "Genetic algorithms with dynamic niche sharing for multimodal function optimization," in *Proceedings of the IEEE International Conference on Evolutionary Computation*, 791, p. 786, IEEE, Nagoya, Japan, May 1996.

[38] V. Narayan and G. Subbarayan, "An optimal feature subset selection using GA for leaf classification," *The International Arab Journal of Information Technology*, vol. 11, no. 5, pp. 447–451, 2014.

[39] D. Napoleon and S. Pavalakodi, "A new method for dimensionality reduction using K-means clustering algorithm for high dimensional data set," *International Journal of Computer Applications*, vol. 13, no. 7, pp. 41–46, 2011.