Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 154 (2019) 109–116

Procedia Computer Science

www.elsevier.com/locate/procedia

8th International Congress of Information and Communication Technology, ICICT 2019

Application in Disease Classification based on KPCA-IBA-LSSVM

Jin Long Jiang [a], Shao Yi Li [a], Ming Liang Liao [b], Yan Jiang [b] *

[a]*Nanchang Institute of Science & Technology, Nanchang 330108, China*
[b]*Tongfang Electronic Technology Corporation, Jiujiang 332005, China*

Abstract

Data mining technology has important clinical significance for disease classification and prevention. In order to improve the performance of the model and the accuracy of disease classification, this paper proposes KPCA-IBA-LSSVM model. In view of the high dimensionality and nonlinearity of medical data, KPCA is used to reduce dimension. BA algorithm is used to optimize the parameters of LSSVM. At the same time, BA algorithm is easy to fall into local extreme and premature convergence. So this paper improves BA algorithm from three aspects. Finally, in order to verify the validity of the algorithm, this paper uses Breast Cancer, Statlog (Heart) and Heart Disease datasets from UCI machine learning database to validate the model. The simulation results show that the model has achieved better classification accuracy, and the model can also be used for classification and prediction of other diseases. The method proves to have certain feasibility and promotion.

Keywords: Data mining; Bat Algorithm(BA), Kernelized Principal Component Analysis (KPCA), Least Square Support Vector Machine(LSSVM);

1. Introduction

As the wide application of electronic medical records and the sharing of information, the hospital database become more and more huge and the key of medical researches. Medical data mining presents the clinical manifestations, development rules and interrelationships among various diseases, which will be of great value to

* Corresponding author. Tel.: +0-86-15951606950.
E-mail address: 420179872@qq.com

10.1016/j.procs.2019.06.017

clinical, teaching and scientific research. In particular, it is of important practical significance for disease prediction, for it will greatly improve the prevention of disease and reduce the incidence of new diseases. Data mining technology is widely used in the medical field, but the model is relatively single and some parameters are set manually. Besides, medical data has the characteristics of wide dimensions, noise, strong coupling and non-linearity, which can not optimize the performance of the model. Therefore, the KPCA-IBA-LSSVM model is proposed to classify diseases.

Kernelized principal component analysis is a nonlinear dimensionality reduction method based on kernel technique. Least square support vector machine is an improved algorithm of SVM. It transforms the quadratic programming problem of SVM into the problem of linear equations, which makes the problems easier to solve. In addition, the performance of the classifier is closely related to the selection of parameters. In this study, the Bat algorithm is used to optimize the parameters. In view of the shortcomings of BA algorithm, such as, likeliness to fall into local extreme value and premature convergence, the bat algorithm is improved from three aspects and an improved Bat algorithm is proposed.

2. Kernelized Principal Component Analysis(KPCA)

Generally, PCA is a dimensionality reduction algorithm for linear data, but the medical data has the characteristics of high dimensionality, multi-noise, strong coupling and nonlinearity, which requires nonlinear mapping to find appropriate low-dimensional embedding. KPCA, a nonlinear dimensionality reduction algorithm, can just solve this problem. The most important is that the KPCA is easy to implement only with the solution of the eigenvalue problem. Besides, it can deal with a variety of nonlinear problems[1-2]. Therefore, it has been widely applied.

3. Bat Algorithm(BA)

Bat algorithm is designed based on the behavior of bats. The echolocation feature of bats is used for searching prey. As it approaches the prey, the bat's pulse emission rate increase and the loudness decrease. In Bat algorithm, n bat individuals update the frequency, velocity and the value of position in the process of flight. The whole algorithm includes two sections, the global updating and local updating. the global updating are shown as follows[3]:

$$f_i = f_{min} + \beta\left(f_{max} - f_{min}\right) \tag{1}$$

$$v_i^t = v_i^{t-1} + f_i\left(x_i^{t-1} - x^*\right) \tag{2}$$

$$x_i^t = x_i^{t-1} + v_i^t \tag{3}$$

Where indicate the varying frequency in the current, β is a random constant and the rang is, $\beta \in [0,1]$ is the current velocity for the *i*th bat, x_i^t is the current position for the ith bat, x^* is the best position for the *i*th bat.

The bat's pulse emission rate and the loudness is updated continuously in the process of local search. The local updating are shown as follows:

$$r_i^t = R_0\left[1 - e^{-\gamma(t-1)}\right] \tag{4}$$

$$A_i^t = \alpha A_i^{t-1} \tag{5}$$

Where r_i^t is the change rate of pulse rate for *i*th bat at t moment, R_0 is the biggest change rate of pulse rate, γ the coefficient parameter of pulse rate which is a constant larger than zero, A_i^t is the loudness of the *i*th bat at *t* moment and A_i^{t-1} is the loudness of the ith bat at *t*-1 moment, α is the coefficient parameter of loudness wave, and its range is [0,1].

During the local searching, the random walk will be used to adjust the position. The equation is shown as follows[4].

$$x_{inew} = x_{iold} + \varepsilon A_i^t \tag{6}$$

where ε is a random number and the rang is[-1,1].

3.1. Improved bat algorithm(IBA)

3.1.1 Improved method

(1) Population initialization based on chaotic and reverse learning strategies

The initialization of population will have great effects on the optimization of the algorithm .The traditional bat algorithm adopts random initialization, which will make the population unevenly distributed. The algorithm is easy to converge to local extreme. Chaos is a unique phenomenon in the nonlinear dynamics, which is characterized by inherent randomness and ergodicity. It can make up for the shortcomings of the random population and improve the quality of population. The expression of chaotic mapping is shown as follows:

$$H_k = \sin\left(\pi H_{k-1}\right), k = 1,\mathrm{L}\ , K \tag{7}$$

Where k is the iteration number and the *K* is the maximum iteration number.

The reverse learning mechanism produces the current individuals and the corresponding reverse individuals. The corresponding opposite individual is generated according to the following formula.

$$ox_{ij} = x_{minj} + x_{maxj} - x_{ij}, i = 1,2,\mathrm{L}\ , N, j = 1,2,\mathrm{L}\ , D \tag{8}$$

Base on the advantage of two methods, This paper proposed a population initialization method based on chaotic reverse learning strategy[5].

(2) Linear decreasing weight

The larger weight is advantageous to jump out the local minimum point and global search, while the smaller weight is helpful to carry out accurate local search in current area and the convergence of the algorithm. Therefore, this paper adopts the linear decreasing weight method, that is, the weight is linear decreased form large to small in order. The weight formula is shown as follows[6]:

$$w = w_{max} - \frac{t\left(w_{max} - w_{min}\right)}{t_{max}} \tag{9}$$

Where *w* is the weight and the range is[w_{min} , w_{max}], t is the number of iteration, t_{max} is the maximum iteration number .The formula update of individual displacement and speed is shown as follows.

$$v_i^t = wv_i^{t-1} + f_i\left(x_i^{t-1} - x^*\right) \tag{10}$$

$$x_i^t = wx_i^{t-1} + v_i^t \tag{11}$$

(3) Natural selection

In order to improve the speed of convergence, natural selection mechanisms are referred to. In each iteration, the particle swarm is sorted according to the fitness value. The worst half of the fitness value is replaced by the best half and the historical optimal value of each individual is kept[7].

3.1.2 The Simulation test and Analysis

In order to verify the performance of the IBA algorithm, the BA and IBA algorithms were respectively tested with four test functions. These functions have many local optimal values and only one global optimal value which is difficult to search. So, the performance of the IBA could be verified by these functions. To prevent accidental errors, the BA and IBA run twenty times independently. The algorithm parameters are set as follows. The population size is 30, The maximum iteration number is 100, The range of frequency is[-1,1], The range of maximum pulse rate A_i is[1,2], The maximum loudness r_i is a random$\in$[0,1], The coefficient parameter of pulse rate γ is 0.9, The coefficient parameter of loudness wave α is 0.9, The maximum chaotic iteration number K is 400, the descending linearly weight w_1, w_{max} and w_{min} is 0.5, 0.5 and 0.1 respectively. The figure1 are the curves of fitness value. And table 2 are the optimization results of BA and IBA algorithm. The test functions are shown as follows.

Table.1 Standard test functions

Function name	Formula	Searching space	The optimal value in theory
Griewank	$f(x)=\sum_{i=1}^{n}\frac{x_i^2}{4000}-\prod_{i=1}^{n}\cos\left(\frac{x_i}{\sqrt{i}}\right)+1$	$x_i\in[-600,600]$	$f(0,\mathrm{L},0)=0$
Rastrigin	$f(x)=\sum_{i=1}^{n}\left[x_i^2-10\cos(2\pi x_i)+10\right]$	$x_i\in[-5.12,5.12]$	$f(0,\mathrm{L},0)=0$
Rosenbrock	$f(x)=\sum_{i=1}^{n}\left[100\left(x_i^2-x_{i+1}\right)^2+(x_i-1)^2\right]$	$x_i\in[-2.048,2.048]$	$f(0,\mathrm{L},0)=0$
Sphere	$f(x)=\sum_{i=1}^{n}x_i^2$	$x_i\in[-10,10]$	$f(0,\mathrm{L},0)=0$

Table.2 The optimization results of test functions

Function	Algorithm	Optimal	Worst	Average	standard deviation
Griewank	BA	1.316e-5	9.007e-5	5.14e-5	3.689e-5
	IBA	1.034e-7	6.891e-9	7.250e-8	1.263e-7
Rastrigin	BA	0.003	0.293	0.580	0.749
	IBA	1.350e-5	8.877e-6	5.020e-5	8.254e-5
Rosenbrock	BA	-404.000	-374.052	-395.44	7.471
	IBA	-404.000	-403.991	-403.994	0.006
Sphere	BA	1.088e-4	9.170e-5	1.370e-4	1.160e-4
	IBA	1.096e-7	9.615e-9	1.200e-7	1.413e-7

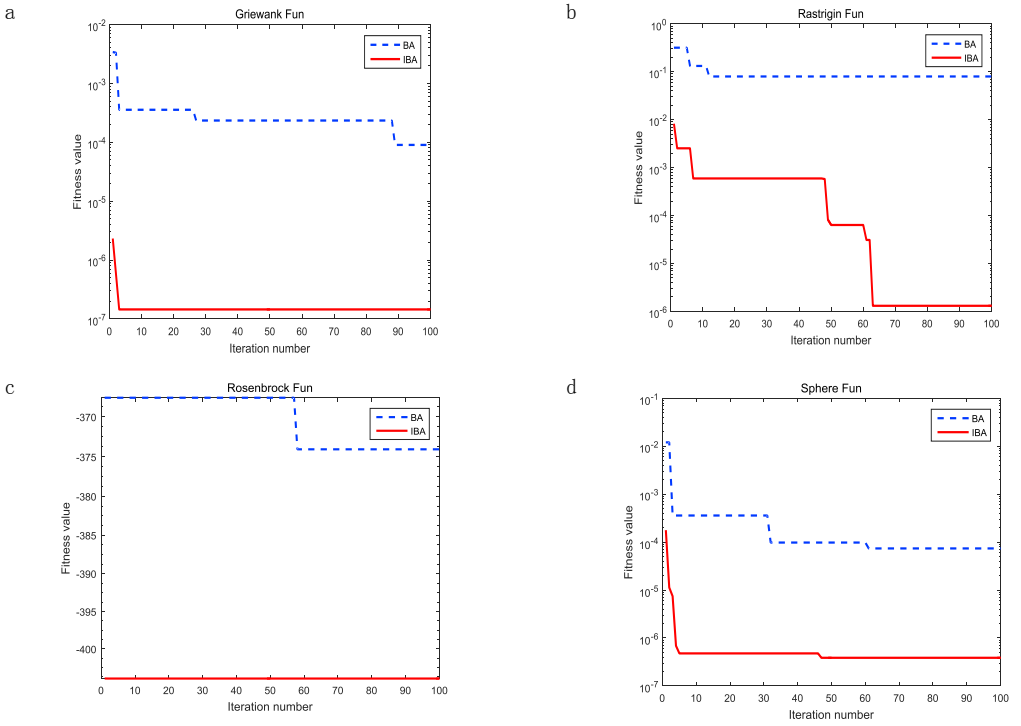


Fig.1 .Comparison of optimization curves of different functions. (a) Comparison of optimization curves of Griewank; (b) Comparison of optimization curves of Rastrigin; (c) Comparison of optimization curves of Rosenbrock; (d) Comparison of optimization curves of Sphere.

Comparing the fitness curves of BA and IBA, we can see the BA algorithm has the characteristics of low individual quality, falling into premature easily and low optimization precision. However, the improved BA algorithm shows better performance both in the early and late stage. In order to prevent accidental error, the BA and IBA run 20 times respectively. The results show that IBA superior to BA in the optimal value, mean and standard deviation. Where the mean reflects the average precision under the maximum number of iterations and the standard deviation reflects the stability of the algorithm. In conclusion, compared with the bat algorithm, the improved bat algorithm greatly improves the speed of convergence, accuracy and stability .The improved bat algorithm can be used to optimize parameters and improve the performance of the model.

4. Least Square Support Vector Machine(LSSVM)

LSSVM is the improved algorithm of SVM proposed by Suykens & Vandewalle. SVM is mainly used to solve the quadratic programming problem, which makes the computation more complex and time-consuming. In order to overcome the above shortcomings, LSSVM replaced inequality constraints in SVM with equality constraints, which changed the quadratic programming problem into liner equations problem, which greatly simplified the problem and accelerated the calculation of the model[8]. In regression case LSSVM is written as[9-10]:

$$\min F = \frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w} + \frac{1}{2}\gamma\sum_{k=1}^{N} e_k^2 \tag{12}$$

$$\mathrm{s.t.}\ \boldsymbol{y_k} = \boldsymbol{w}^{\mathrm{T}}\varphi\left(\boldsymbol{x_k}\right) + b + e_k \tag{13}$$

Where γ is a regularization parameter , e_k is the error term, $\boldsymbol{w}$ is the support vector , b is the threshold. And $\boldsymbol{x}_k$, $\boldsymbol{y}_k$ is the input and output respectively.

The LSSVM model is shown as below:

$$f(\boldsymbol{x}) = \sum_{k=1}^{N} \alpha_k k(\boldsymbol{x}, \boldsymbol{x}_k) + b \tag{14}$$

Where α_k are Lagrange multipliers, $k(\boldsymbol{x}, \boldsymbol{x}_k)$ is the kernel function. There are many types of kernel functions. And the RBF-kernel is the most widely used. The equation is shown as below. In order to make the model achieve to the best performance, we need to select the appropriate kernel function.

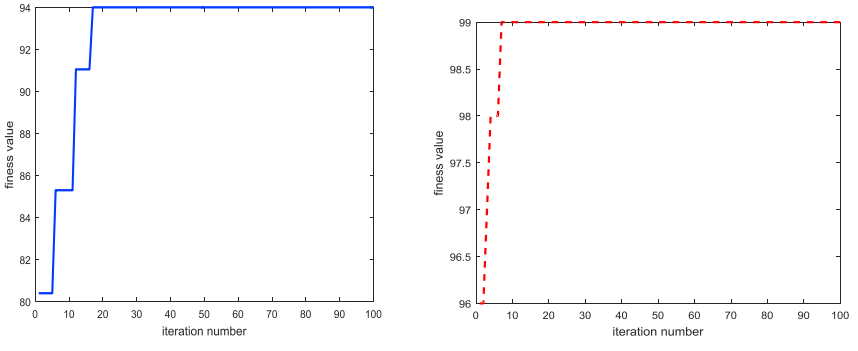
$$k(\boldsymbol{x}, \boldsymbol{x}_k) = \exp\left(-\left\|\boldsymbol{x}\text{-}\boldsymbol{x}_k\right\|^2 / 2\sigma^2\right) \tag{15}$$

4.1. Simulation Results

Breast Cancer, Statlog (Heart), Heart Disease were selected from UCI machine learning database as samples. The algorithm parameters are set as follows: The population size is 30;The maximum iteration number is 100; The range of frequency is [-1,1]; The range of maximum pulse rate A_i is [1,2]; The maximum loudness r_i is a random $\in$ [0,1].The coefficient parameter of pulse rate γ is 0.9. The coefficient parameter of loudness wave α is 0.9, The maximum chaotic iteration number K is 400. the descending linearly weight w_1, w_{max} and w_{min} is 0.5, 0.5 and 0.1 respectively. LSSVM chooses the RBF as kernel function, and the solution space is 2 dimension, which γ represents the regularization parameters and δ^2 represents RBF kernel parameters. The γ range of values is [0.1, 1000], and the δ^2 range of values is [0.01,100].

Table.3 Information of data sets

Data set	Sample size	Dimension	Category
Breast Cancer	569	31	2
Statlog(Heart)	270	13	2
Heart Disease	303	14	2



a b

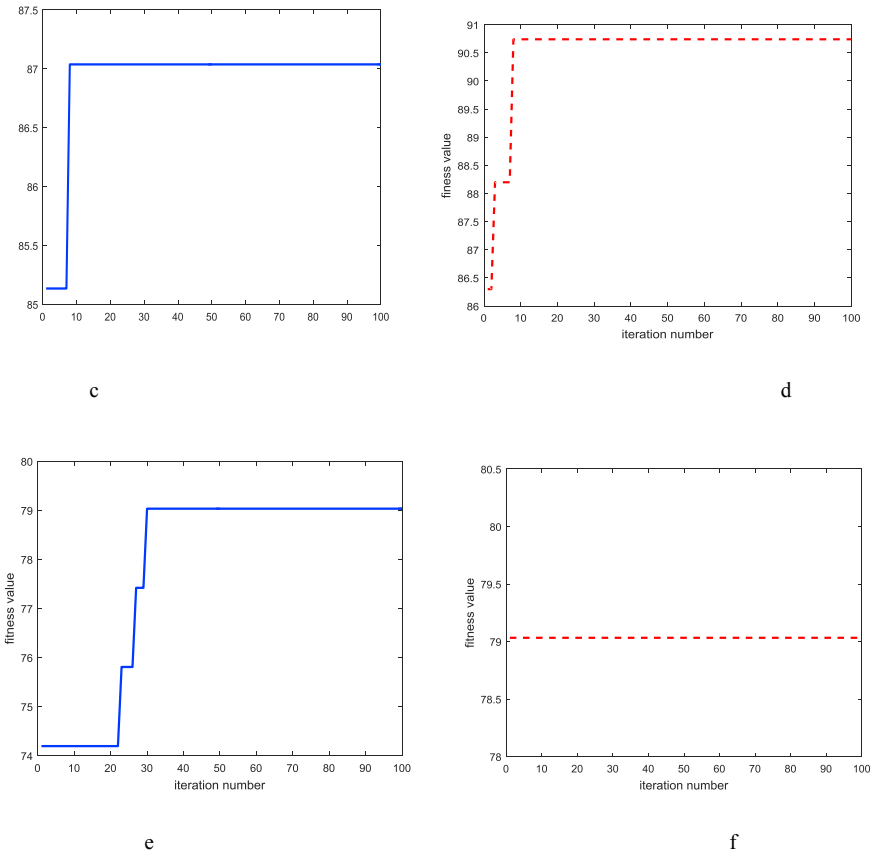


Fig.2 .Fitness curve of different data sets.(a) The fitness curve of breast cancer with KPCA-BA-LSSVM model; (b)The fitness curve of breast cancer with KPCA-IBA-LSSVM model; (c) The fitness curve of Statlog(Heart) with KPCA-BA-LSSVM model; (d)The fitness curve of Statlog(Heart) with KPCA-IBA-LSSVM model ;(e) The fitness curve of heart disease with KPCA-BA-LSSVM model.(f)The fitness curve of heart disease with KPCA-IBA-LSSVM model.

Table. 4 Classification accuracy comparison of different data sets

Data set	algorithm	Classification accuracy (%)
Breast Cancer	KPCA-BA-LSSVM	94
	KPCA-IBA-LSSVM	99
Statlog	KPCA-BA-LSSVM	87.0370
	KPCA-IBA-LSSVM	90.7407
Heart Disease	KPCA-BA-LSSVM	79.0323
	KPCA-IBA-LSSVM	79.0323

5. Summary and Prospect

The three datasets are classified by KPCA-BA-LSSVM and KPCA-IBA-LSSVM respectively. The simulation results are shown in Figure 2. The fitness curves of Breast Cancer and Statlog (Heart) show that the model optimized by IBA is better than BA at the beginning of iteration. It shows that the individual quality of KPCA-IBA-LSSVM model is better at the early evolution, the convergence speed of KPCA-IBA-LSSVM model is faster, and its classification is more accurate in the late stage. The Heart disease fitness curve shows that although the final

optimization results of the two models are the same, the model optimized by BA shows obvious inflection point before 35 iterations, and the model falls into local optimum, while the model optimized by IBA has found global optimum in the first generation. The above shows that the latter is more efficient. Tab 3 shows that the three data sets optimized by IBA has achieved better classification results and can be used to classify other diseases. This model has a certain prospect in popularization. At present, there are many kinds of swarm intelligence optimization algorithms. So other optimization algorithms will be used to optimize the model.

References

1. Navi M, Meskin N, Davoodi M. Sensor fault detection and isolation of an industrial gas turbine using partial adaptive KPCA[J]. Journal of Process Control, 2018, 64:37-48.
2. Huang J, Yan X. Related and independent variable fault detection based on KPCA and SVDD[J]. Journal of Process Control, 2016, 39:88-99.
3. Wu Z, Yu D. Application of Improved Bat Algorithm for Solar PV Maximum Power Point Tracking under Partially Shaded Condition[J]. Applied Soft Computing, 62(2018) 101-109.
4. Kora P, Kalva S R. Improved Bat algorithm for the detection of myocardial infarction[J]. Springerplus, 2015, 4(1):1-18.
5. Gao Weifeng, Liu Sanyang, Jiao Hehua, et al.Particle swarm optimization with artificial bee colony search operator[J]. Control and decision making, 2012, 27(6):833-838.
6. Alfi A, Modares H. System identification and control using adaptive particle swarm optimization[J]. Applied Mathematical Modelling, 2011, 35(3):1210-1221.
7. Zhang Jiannan, Liu Yian, Wang Gang. Path planning of UAV Based on particle swarm optimization algorithm[J]. Sensors and Microsystems, 2017, 36(3):58-61.
8. Yarveicy H, Ghiasi M M. Modeling of gas hydrate phase equilibria: Extremely randomized trees and LSSVM approaches[J]. Journal of Molecular Liquids, 2017, 243:553-541.
9. Baghban A. Prediction viscosity of ionic liquids using a hybrid LSSVM and group contribution method[J]. Journal of Molecular Liquids, 2017, 236:452-464.
10. Dong R, Xu J, Lin B. ROI-based study on impact factors of distributed PV projects by LSSVM-PSO[J]. Energy, 2017, 124:336-349.