Variables Screening Method Based on the Algorithm of Combining Fruit Fly Optimization Algorithm and RBF Neural Network

Fuqiang Xu,Youtian Tao
Department of Mathematics
Chaohu College
Chaohu China
13966322177@139.com

Abstract—**The form of fruit fly optimization algorithm (FOA) is easy to learn and has the characteristics of quick convergence and not readily dropping into local optimum. This paper presents the optimization of RBF neural network by means of FOA and establishment of network model, adopting it with the combination of the evaluation of the mean impact value(MIV) to select variables. The validity of this model is tested by two actual examples, furthermore, it is simpler to learn, more stable and practical.**

Keywords-FOA; RBF neural network; Parameter optimization; MIV; Variables screen

I. Introduction

We have statistics of a large number of experimental data in many scientific experiments such as Near Infrared Spectral data[1, 2] and Atlas data[3, 4]. Our aim is to find a bivariable function based on such a large number of experimental data. But this kind of function is often highly uncertain, nonlinear dynamic model. When we perform on the data regression analysis, this requires choosing appropriate independent variables to establish the independent variables on the dependent variables regression model. Generally, experiments often get more variables, some variables affecting the results may be smaller or no influence at all, even some variable acquisition need to pay a large cost. If drawing unimportant variables into model, we can reduce the precision of the model, but can not reach the ideal result[5]. At the same time, a large number of variables may also exist in multicollinearity. Therefore, the independent variable screening before modeling is very necessary[6]. Because the fruit fly optimization algorithm has concise form, is easy to learn, and have fault tolerant ability, besides algorithm realizes time shorter, and the iterative optimization is difficult to fall into the local extreme value. And radiate basis function (RBF) neural network's structure is simple, training concise and fasting speed of convergence by learning, can approximate any nonlinear function, having a "local perception field" reputation. For this reason, this paper puts forward a method of making use of the fruit flies optimization algorithm to optimize RBF neural network (FOA-RBF algorithm) using for variable selection.

II. About Fruit Fly Optimization Algorithm

Fruit Fly Optimization Algorithm, refers to FOA,is the latest of evolutionary computation technology, put forward by Pan Wenchao of Taiwan in 2011. FOA algorithm is a new method of swarm intelligence of global optimization performance, basing on fruit flies foraging behavior. Because fruit flies in the sensory and perception is superior to other species, especially in the vision and the sense of smell. Fruit flies olfactory organs can be a very good collection of various air smell, can even smell food source about 40 km away. When flying close to the food, Fruit flies use the keen vision and companions gathers to the position to find the food position, then fly to the direction[7].

The algorithm's basic steps:

Step 1, random initial the position of the fruit flies group(X-axis,Y-axis).

Step 2, endow fruit flies individual's random direction（Random Value）and position (X，Y) rely on food smell.

X= X-axis+ Random Value
Y= Y-axis+ Random Value

Step 3, calculate the distance of fruit flies individual and origin of coordinates(Dist)
, and calculate the taste concentration decision value S, which for the distance of the reciprocal.

$$\text{Dist}=\sqrt{X^2+Y^2}\text{;S=}\frac{1}{Dist}.$$

Step 4, put the taste concentration decision value (S) substitution taste concentration decision Function (Function), find out the location of the individual's taste concentration (Smell). Smell=Function(S)

Step 5, repeat step 2 to step 4, calculate each individual's taste concentration, and find out the highest (lowest) taste concentration of fruit fly.

Step 6, save the best taste of fruit fly's value S and position (X, Y). At this point, the group of fruit flies flies to this position.

Step 7, turn into the iterative optimization, repeat step 2 to step 5, to see if the taste value S is better than previous iterative taste value S, if so, turn to step 6.

III. The Radial Basis Neural Network Introduction

The radial basis neural network is also called RBF neural network, belongs to the forward type neural network. The

Published by Atlantis Press, Paris, France.

network's structure is similar to the multilayer feed forward network, only has a hidden layer of three layer forward network. The first layer is input layer, which composed by signal node; The second is hidden layer, the layer's node decided by apparent problem, the layer neuron transformation function that radial basis function is the local response of the Gaussian function; The third layer is output layer. The Figure. 1 is structure of rbf neural network.

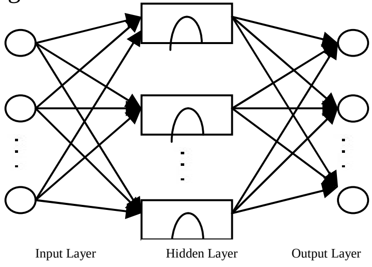


Figure 1. The structure of rbf neural network

RBF network use radial basis as hidden unit "base" to form a hidden layer, which transforms the input vector, the low dimensional model of input data converse into a high dimensional space inside, this will change a low dimensional space in linear inseparable problem into a high dimensional space in linear separable, and the problems are solved.

RBF neural network has a variety of learning methods. This paper is based on the self-organizing selection center learning method. This learning process is divided into two steps: the first step is self-organization learning stage, which solves hidden layers' center and variance of the basis function; The second step solve weights between hidden layer and output layer. In the training process, determining the number of hidden layer neurons is a key issue, the basic training principle starts from zero neurons, through checking output error whether meet the requirements, or else using in each cycle by automatic increasing neurons, so that network generates maximum error corresponding input vector as weight vector, produce a new neurons, then check the new network error, repeat this process until error requirements or maximum number of hidden layer neurons. Thus it can be seen, RBF neural network has simple structure, adaptive, output and initial weights of the independent characteristics, and the independent learning convergence speed, can approximate any nonlinear function [9].

IV. FOA-RBF Algorithm Used to Implement Variable Screening

Parameter SPREAD is radial basis function's expansion speed. Building good RBF neural network, SPREAD value directly affect network fitting (prediction) accuracy when the network learning training, choice reasonable SPREAD value is very important. The greater SPREAD value, can make the radial basis neurons covered to input vector space which has response, but also do not need all the radial basis neurons has response, as long as part of the response is enough, and too big SPREAD value can also lead to the difficulty of computing. In the design of network often try different SPREAD value, has certain subjectivity and uncertainty, is not easy to get the optimal model [10].

In order to seek the optimal model, RBF neural network can be drawn into FOA algorithm iterative optimization process, taste concentration decision value(S=1/Dist>0) as SPREAD value directly, the network absolute value of prediction error's sum as tasting concentration decision function Smell value.

In order to prevent the RBF network over fitting and influence network promotion ability, we divide it into two groups of experimental data cross validation, calculate the each network's training error, find that iterations and the corresponding Smell value and SPREAD value when network absolute value of prediction error's sum achieved minimum, so as to find the best SPREAD value. Then use this SPREAD value and all input and output samples to establish optimal RBF network. After that we draw the mean influence value (MIV) into the optimal RBF neural network for main variable screening.

MIV is one of the important indexes which is used to evaluate the affects of each independent variable to the dependent variables. MIV led into neural network is an index, which can reflect the network weight matrix, changes of the input neurons, and evaluation of the size of the output neurons influence. At present, the MIV is considered one of the best indexes in the neural network evaluate variables in the correlation. MIV's symbol stands for input to output related direction, its absolute size represents the relative importance of the influence.

After training in established optimal RBF network, each independent variable of the training samples $\pm 10\%$ in the original basis, constitute two new training sample K1 and K2, I set them as the new training sample substitution into established optimal RBF network ,then simulate them respectively, get two output respectively S1 and S2, calculate the difference value of S1 and S2, treat the outcome as a changed independent variable to output variable value(IV) ,finally according to the training sample size calculate IV's average value, get the MIV conclusion that independent variables to output variable (dependent variable). According this method, calculating each independent variable MIV, arranging each independent variable order according to the absolute value of the MIV, getting relative important sequence from each variable to dependent variable, so we can realize the variable screening[5].please see the following Figure. 2.

V. Application Examples

A. Data preparation

1). we generate five hundred x1, x2 and x3 value randomly and respectively in region [-2, 2], produce Y1 value in according with the three dimensional nonlinear function

$$Y_1 = x_1^2 + 2x_2^2 + 5x_3^2 + 2x_1x_2 + 6x_2x_3 + 2x_1x_3 .$$

We still randomly generated five hundred x4 and x5 value as interference variables, and then store the data for Data1.

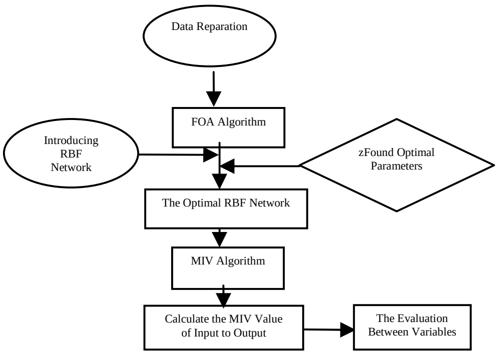

Data Reparation
FOA Algorithm
Introducing RBF Network
zFound Optimal Parameters
The Optimal RBF Network
MIV Algorithm
Calculate the MIV Value of Input to Output
The Evaluation Between Variables

Figure 2. Variable screening process of foa-rbf algorithm

2) .we generate five hundred x1 and x2 value randomly and respectively in the region[-4, 4], produce Y1 value which are in according with two dimensional nonlinear functions $Y_2 = (x_1 + x_2{}^2 + 2x_2)e^{x_1+x_2}$. We still random generated five hundred x3, x4 and x5 value as interference variables, then store the data for Data2.

x1, x2, x3, x4, x5 and Y1 are regarded as the training sample. xi, i = 1, 2, 3, 4, is regarded as network input (independent variables). Y1 is regarded as the network output (dependent variable). The FOA - RBF algorithm combining with MIV method to screen out the output results are mainly on the effects of the variables. The experiment selects the suitable variables which have main affection to output Y through the FOA-RBF algorithm combining MIV method.

B. Constructing model

We have 500 sample data normalization processing. The results are divided into two groups, each has 200 sample data as a set of training samples. Fruit flies optimization algorithm's iteration number are set to 100, population scale are set to 10, the 'I' fly's initial position (X, Y) set to X(i)=X_axis+2*rand()-1; X-axis = rand(); Y(i) =Y_axis+2*rand()-1; Y-axis= rand(). Taste concentration decision value is S(i)=1/sqrt(D(i)); SPREAD value is SPREAD = S (i).By using the SPREAD value and 'newrb()' function create a approximation radial basis RBF network, , one group of the two for learning, another is used to predict (iterative when cross validation), in the network's learning process we try to increase the number of hidden layer neurons constantly until the network output error can satisfy the preset value so far.

We consider the prediction error's absolute value sum as taste density smell value. And find the minimum value of the population Smell, record the corresponding iterations, taste density (SPREAD value) and the fruit fly individual position. Then we find the best SPREAD value during the iterative optimization process, and establish the optimal RBF network. Finally, we count the MIV value of each input variables to output variable by MIV algorithm.

C. Experimental results and analysis

FOA - RBF algorithm for data Data1's operation result is shown from Figure. 3 to Figure. 5. Algorithm's running time is 54.217590 s.

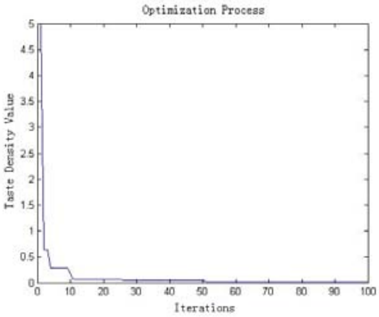


Figure 3. Foa - rbf algorithm's iterative optimization process of data 'data1'

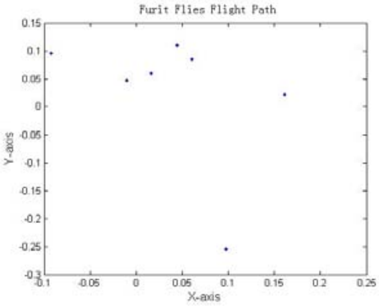
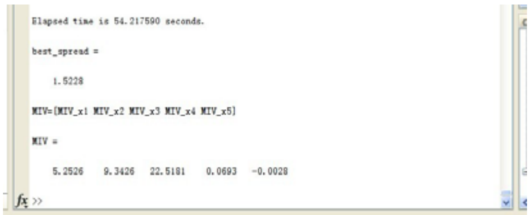


Figure 4. The flies' optimal flight path of data 'data1'



```
Elapsed time is 54.217590 seconds.

best_spread =

    1.5228

MIV=[MIV_x1 MIV_x2 MIV_x3 MIV_x4 MIV_x5]

MIV =

    5.2526    9.3426   22.5181    0.0693   -0.0028
```

Figure 5. Foa - rbf algorithm's results of data 'data1'

The best SPREAD value is 1.5228. The corresponding fruit flies' coordinates of this group is (0.0476, 0.1159). MIV_Xi value is in accordance with MIV value and the independent variable xi to network's output Y1,which is 5.2526, 9.3426,22.5181, 0.0693 and -0.0028 respectively. MIV's symbol stands for the related direction of input to output; its absolute size represents the relative importance of the influence. The algorithm continuously operate ten times (sample data is randomly generated each time). All results see in table 1.

TABLE I. THE RESULTS OF CONTINUOUS OPERATION 10 TIMES ABOUT 'DATA1' BY FOA-RBF ALGORITHM

The serial number	**Algorithm running time（s）**	**The best SPREAD value**	**The independent variable's MIV value（MIV_x1，…，MIV_x5）**
1	54.21759	1.5228	5.2526,9.3426,22.5181,0.06

	0		93, -0.0028
2	52.264266	0.8634	5.6064,11.4374,24.2477,1.1267, 1.5267
3	52.966207	1.1064	5.0192,11.3304,26.7134,0.1542, 0.1904
4	52.226382	1.0615	4.1845,10.7414,25.8335,-1.0246, -0.9056
5	52.059187	1.0535	4.9435,9.4303,27.4075,0.2438, 0.2521
6	52.390677	1.0640	5.7909,10.5295,24.2180,0.7234,0.6983
7	52.384335	0.9812	6.2564,11.5141,25.6123,0.6376,0.8288
8	52.242513	0.9162	4.7596, 10.5498、23.7379, 0.2212, 0.1021
9	52.553951	1.2756	5.2684,9.9638,24.0290,0.2586, 0.2342
10	52.456818	1.0992	5.2056,10.0444,25.3855,0.2880,0.2151

This shows that the first three variables' MIV value of x1, x2, and x3 are bigger. These values are calculated by x1, x2, and x3, has nothing to do with x4 and x5. Therefore, the results of the main effect to dependent variable Y1 which screened out by this algorithm are real conditions consistent.

Also, the operation results for data 'Data2' seen from Figure. 6 to Figure. 8 by use of FOA - RBF algorithm. Algorithm running time is 52.070761s. The best SPREAD value is 0.3585, the group of fruit flies position coordinates for (2.2335, 1.2353) eventually. MIV_ Xi's value is 127.6204, 169.4325, 41.3302, 37.2808, 19.3070. The algorithm also continuous ten times (sample data is random generated each time), the results see table 2.

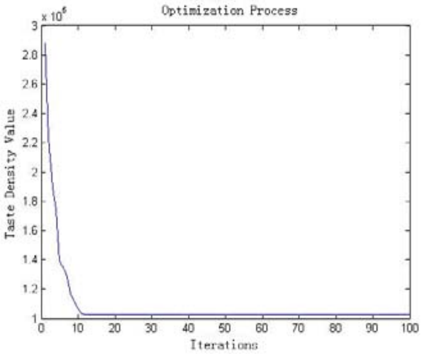


Figure 6. Foa - rbf algorithm's iterative optimization process of data 'data2'

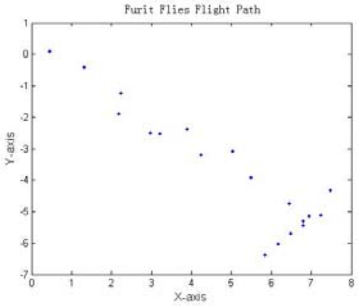
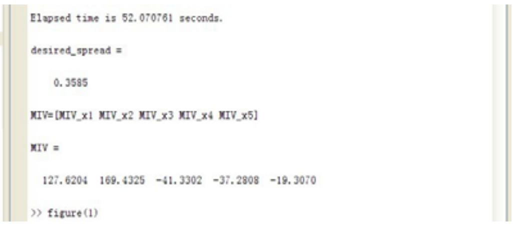


Figure 7. The flies' optimal flight path of data 'data2'



```
Elapsed time is 52.070761 seconds.

desired_spread =

    0.3585

MIV=[MIV_x1 MIV_x2 MIV_x3 MIV_x4 MIV_x5]

MIV =

  127.6204  169.4325  -41.3302  -37.2808  -19.3070

>> figure(1)
```

Figure 8. Foa - rbf algorithm's results of data 'data2'

TABLE II. THE RESULTS OF CONTINUOUS OPERATION 10 TIMES ABOUT 'DATA2' BY FOA-RBF ALGORITHM

The serial number	**Algorithm running time（s）**	**The best SPREAD value**	**The independent variable's MIV value（MIV_x1，…，MIV_x5）**
1	52.070761	0.3585	127.6204,169.4325,-41.3302, -37.2808,-19.3070
2	56.912273	0.4949	142.1258,156.5311,-65.8890, -76.5729,-59.8165
3	52.955942	0.4387	168.1514,196.9594,16.5550, -28.6309,-80.8060
4	58.133814	0.2383	71.5886,115.8796,-2.9341, -7.4295,-14.3774
5	57.300290	0.7439	416.9257,531.0809,130.2046, 97.7198,164.7629
6	56.399827	0.4410	343.6559、395.7168、 24.1714、120.3018、 45.8121
7	56.095968	0.4580	142.7104,293.7982,-18.7317, -20.4932,-25.2040
8	58.383318	0.3587	134.4711,205.6932,-56.1161, -45.3180,-15.7042
9	58.026312	0.2935	96.4039,90.9052,-9.7226, -32.3673,-39.4657
10	55.433284	1.4905	425.0835,601.4664,-21.2851, 46.8663,63.6262

Table 2 shows that the first two variables' MIV value of x1 and x2 are bigger. These values are calculated by x1 and x2, which has nothing to do with x3, x4 and x5. Therefore, the results of the main effect to dependent variable Y2 which

Published by Atlantis Press, Paris, France.

screened out by this algorithm are also consistently real conditions.

VI. SUMMARIZE

Compared with the other neural network such as BP, ELman etc, For RBF net has the below advantage: small quantity adjustable parameter, simple structure, self-addapting, the loading results has no business with original value etc. The Training results of RBF net has much relations with the value of only parameter Spread, this article using FOA to optimizing RBF neural network, get the best Spread value by FOA Iterative optimization, then build the best RBF net. Finally, combine with the average impact value evaluation, get the MIV(the value of importance that independent variable to dependent variable). Try to bring in 2 or 3 disturbing variable by two example, FOA-RBF algorithm can screen successfully, the percentage merely match 100% . Which means it is workable to screen master variable for FOA imization algorithm by optimizing the RBF net and combine with the MIV algorithm. Compare with least square method[1], Combined with the genetic algorithm of least squares[3,4] etc, in addition imization algorithm is more simple, and easily learned, Convergence fast and not easy to trap in the local extremum , which make FOA-RBF has better stability and usability.

In addition, the Optimization ability of FOA imization algorithm restricted by group size, iterations , Progress value etc, but has small effection after writer's many times testing, Can not show it one by one for article length reason. So, FOA-RBF algorithm is antother good way to realize the variable screen.

REFERENCES

[1] WU Rui-mei1,ZHAO Jie-wen,CHEN Quan-sheng and HUANG Xing-yi. Determination of Taste Quality of Green Tea Using FT-NIR Spectroscopy and Variable Selection Methods,Spectroscopy and Spectral Analysis,2011,31(7).pp.34-37.

[2] Xu Heng.Statistic methods for variable selection and robust modeling in near infrared spectral analysis. Nankai university: Analytical Chemistry, 2010.

[3] Chu Xiaoli, Yuan Hongfu, Wang Yanbin and Lu Wanzhen.Variable Selection for Partial Least Squares Modeling by Genetic Algorithms.Chinese Journal of Analytieal Chemistry, 2001,29(4).pp.437-442.

[4] Wang Guoqing and Shao Xueguang.A Discrete Wavelet Transform-Genetic Algorithm-Cross Validation Approach for High Ratio Compression and Variable Selection of Near-infrared Spectral Data.Chinese Journal of Analytical Chemistry,2005,33(2).pp.191-194.

[5] XU Fu-Qiang and LIU Xiang-Guo.Variables Screening Methods Based on the Optimization of RBF Neural Network.Computer Systems & Applications,2012，21(3).pp.206-208.

[6] Lin Yan.The PLS Variable Selection Method and Its Application.Xiamen university: Analytical Chemistry, 2007.

[7] PAN Wen-chao.Fruit Fly Optimization Algorithm. Taipei: The Sea Press,2011.pp.11-12.

[8] PAN Wen-chao.Using Fruit Fly Optimization Algorithm Optimized General Regression Neural Network to Construct the Operating Performance of Enterprises Model.Journal of Taiyuan University of Technology(Social Sciences Edition),2011,39(4).pp.1-4.

[9] Ge Zhexue and Sun Zhiqiang.The neural network theory and realize by Matlab r2007. Beijing: Electronic Industry Publish,2007.pp.117-120.

[10] ZHANG Gang-lin and LIU Guang-can. Parameter Optimization of RBF Networks Based on Evolutionary Model. Control Engineering of China, 2010,17(3).pp.67-70.

Published by Atlantis Press, Paris, France.