Study on the PPE Model Based on RAGA to Evaluating the Water Quality

Qiang Fu[1, 2], Wei Zu [3]

1. Doctoral Working Station of Beidahuang Company, Total Bureau of Agricultural in Heilongjiang，Harbin Heilongjiang 150040, China;
2. School of Water Conservancy & Civil Engineering, Northeast Agriculture University, Harbin Heilongjiang 150030, China;
3. School of Agriculture, Northeast Agriculture University, Harbin, Heilongjiang 150030, China

Abstract: This study improved the traditional genetic algorithm, and combined the new method named Real coding based Accelerating Genetic Algorithm (RAGA) with Projection Pursuit Evaluation (PPE) model. The RAGA-PPE model can optimize several parameters at one time. Based on this method, the authors built up a new evaluating model. Through applying the new model to evaluating nutrient states of south lake in Changchun, it gained the good results. Thus, we provided a new method and thought for readers who engage in researching the water quality evaluation. [Nature and Science. 2004;2(4): 34-38].

Key words: Real coding based Accelerating Genetic Algorithm (RAGA); Projection Pursuit Evaluation (PPE); water quality evaluation; model

1. Introduction

Water quality evaluation is to evaluate synthetically for the grade of water quality through building up mathematical model based on some evaluated indexes. Thus, we can provide some scientific gist for management and preventing pollution according to model. At present, there are many methods to evaluate the water quality. Such as gray clustering, fuzzy clustering, artificial nerve nets, (ANN) matter-element model and so on. (Jin, 2000). These models have the disturbance of giving weight by artificial factors and can't distinguish the grade precisely. Because the synthetic evaluation will be determined by many non-linearity indexes, when we build up model to classify and evaluate using traditional method, it is very difficult to find the internal rule. Recently, Friedme put forward a new arithmetic named Projection Pursuit (PP) which is fit for multivariate statistical analysis (Zhang, 2000). The kind of method can solve many non-linearity problems in a certain extent. But it is very difficult to find the best projection direction owing to the complicated space structure of multi-dimension data. Therefor, the author adopts Real coding based Accelerating Genetic Algorithm RAGA) to optimize the projection direction. RAGA can find the best value in the whole scope. Through using RAGA to reducing the dimension number, we can translate high-dimension data into the synthetic projection value in low-dimension sub-space. Thus, we can apply PPE model based RAGA to evaluate the water quality (Jin, 2000; Zhou, 2000).

2. Projection Pursuit Evaluation Model (PPE)

2.1 Brief introduction of PP model

The main characteristics of PP model are as follows. Firstly, PP model can handle the difficulty named dimension disaster, which have been brought by high-dimension data. Secondly, PP model can eliminate the jamming, which are irrespective with data structure. Thirdly, PP model provides a new approach to handle high-dimension problem using one dimension statistics method. Fourthly, PP method can deal with non-linearity problem (Jin, 2000; Zhang. 2000).

2.2 Step of PPE modeling

The step of building up PPE model includes 4 steps as follows (Jin, 2000; Zhang. 2000).

Step 1: Normalizing the evaluation indexes set of each sample. Now, we suppose the sample set is $\{x^*(i,j)|i=1\sim n, j=1\sim p\}$. $x^*(i,j)$ is the index value of j and sample of i. n ——the number of sample. p ——the number of index. In order to eliminate the dimension influence and unite the change scope of each index value, we can adopt the following formulas

to normalize the data.

$$x(i,j)=\frac{x^*(i,j)-x_{min}(j)}{x_{max}(j)-x_{min}(j)} \quad (1—a) \text{ or:}$$

$$x(i,j)=\frac{x_{max}(j)-x^*(i,j)}{x_{max}(j)-x_{min}(j)} \quad (1—b)$$

In formula: $x_{max}(j)$ and $x_{min}(j)$ stand for the max and the min of j index value. $x(i,j)$ is the index list after moralization.

Step 2: Constructing the projection index function $Q(a)$. PP method is to turn p dimension data ($\{x^*(i,j)|j=1\sim p\}$) into one dimension projection value $z(i)$ based on projection direction a.

$a=\{a(1),a(2),a(3),\cdots,a(p)\}$,

$$z(i)=\sum_{j=1}^{p}a(j)x(i,j) \quad (i=1\sim n) \quad (2)$$

Then, we can classify the sample according to one-dimension scatter figure of $z(i)$. In formula (2), a stand for unit length vector.

Thus, the projection index function can be expressed as follows.

$$Q(a)=S_z D_z \quad (3)$$

In formula: S_z ——the standard deviation of $z(i)$, D_z ——the partial density of $z(i)$.

$$S_z=\sqrt{\frac{\sum_{i=1}^{n}(z(i)-E(z))^2}{n-1}} \quad (4)$$

$$D_z=\sum_{i=1}^{n}\sum_{j=1}^{n}(R-r(i,j))\cdot u(R-r(i,j)) \quad (5)$$

In formula (4) and (5), $E(z)$ ——the average value of series $\{z(i)|\,i=1\sim n\}$; R ——the window radius of partial density, commonly, $R=0.1S_z$. $r(i,j)$ ——the distance of sample, $r(i,j)=|z(i)-z(j)|$; $u(t)$ ——a unit jump function, if $t\geq 0$, $u(t)$ =1,if $t<0$, $u(t)$ =0.

Step 3: Optimizing the projection index function. When every indexes value of each sample have been fixed, the projection function $Q(a)$ change only according to projection direction a. Different projection direction reflects different data structure characteristic. The best projection direction is the most likely to discovery some characteristic structure of high-dimension data. So, we can calculate the max of $Q(a)$ to estimate the best project direction.

Function: $Max: Q(a)=S_z\cdot D_z$ (6)

Restricted condition $s.t: \sum_{j=1}^{p}a^2(j)=1$ (7)

Formula (6) and (7) is a complex non-linearity optimization, which take $\{a(j)|\,j=1\sim p\}$ as optimized variable. Traditional method is very difficulty to calculate. Now, we adopt RAGA to handle the kind of problem.

Step 4: Classification. We can put the best projection direction a^* into formula (2), then we can obtain the projection value of each sample dot. Compare $z^*(i)$ with $z^*(j)$. If $z^*(i)$ is closer to $z^*(j)$, that means sample i and j are trend to the same species. If we dispose $z^*(i)$ from big to small, we can obtain the new sample list from good to bad.

3. Real coding based accelerating genetic algorithm (RAGA)

3.1 Brief introduction of GA

Genetic Algorithm has been put forward by Professor Holland in USA. The main operation includes selection, crossover and mutation (Jin, et al. 2000; Zhou, et al. 2000).

3.2 Real coding based accelerating genetic algorithm (RAGA)

The coding mode of traditional GA adopted binary system. But binary system coding mode has many abuses. So, through consulting literature(Jin, 2000), the author put forward a new method named RAGA (Real coding based Accelerating Genetic Algorithm). RAGA includes 8 steps as follows. For example, we want to calculate the following best optimization problem.

$$Max: \text{f(X)}$$
$$\text{s.t.}: a_j \leq x_j \leq b_j$$

Step1: In the scope of $[a_j,b_j]$, we can create N group uniformity distributing random variable $V_i^{(0)}(x_1,x_2,\cdots x_j,\cdots x_p)$. $i=1\sim N$, $j=1\sim p$. N——the group scale. p —— the number of optimized parameter.

Step 2: Calculate the target function value. Putting the original chromosome $V_i^{(0)}$ into target function, we can calculate the corresponding function value $f^{(0)}(V_i^{(0)})$. According to the function value, we dispose the chromosome from big to small. Then, we obtain $V_i^{(1)}$.

Step 3: Calculate the evaluation function based on order expresses as $eval(V)$. The evaluation function gives a probability for each chromosome V. It makes the probability of the chromosome to be selected is fit for the adaptability of other chromosomes. The better the adaptability of chromosome is, the much easier to be selected. Now, if parameter $\alpha \in (0,1)$, the evaluation function based order can be expressed as follows.

$eval(V_i) = \alpha(1-\alpha)^{i-1}$, i = 1,2,…, N

Step 4: Selecting operation. The course of selecting is based on circumratating the bet wheel N times. We can select a new chromosome from each rotation. The bet wheel selects the chromosome according to the adaptability. We obtain a new group $V_i^{(2)}$ after selecting.

Step 5: Crossover operation. Firstly, we define the parameter P_c as the crossover probability. In order to ensure the parent generation group to crossover, we can repeat the process from $i = 1$ to N as follows. Create random number r from [0，1]. If $r < P_c$, we take V_i as parent generation. We use $V_1^{'}, V_2^{'}, \cdots$ to stand for male parent which to be selected. At the same time, we divide the chromosome into random pair based on arithmetic crossing method. That is as follows.

$X = c \cdot V_1^{'} + (1-c) \cdot V_2^{'}$ $\quad Y = (1-c) \cdot V_1^{'} + c \cdot V_2^{'}$

c ——a random number from (0,1).

We can obtain a new group $V_i^{(3)}$ after crossover.

Step 6: Mutation operation. Define the P_m as mutation probability. We select the mutation direction d randomly from R^n. If $V + Md$ isn't feasible, we can make M a random number from 0 to M until the value of $V + Md$ is feasible. M is a enough big number. Then, we can use $X = V + Md$ replace V. After mutation operation, we obtain a new group $V_i^{(4)}$.

Step 7: Evolution iteration. We can obtain the filial generation $V_i^{(4)}$ from step 4 to step 6, and dispose them according to adaptability function value from big to small. Then, the arithmetic comes into the next evolution process. Thus, the above steps have been operated repeatedly until the end.

Step 8: The above seven steps make up of Standard Genetic Arithmetic (SGA). But SGA can't assure the whole astringency. The research indicates that the seeking optimization function of selecting and crossover has wear off along with the iteration times increasing. In practical application, SGA will stop to working when it is far away from the best value, and many individuals are conform or repeated. Enlightening by reference (Jin, 2000), we can adopt the interval of excellence individual during the course of the first and the second iteration as the new interval. Then, the arithmetic comes into step 1, and runs SGA over again to form accelerate running. Thus, the interval of excellence individual will gradually reduce, and the distance is closer to the best dot. The arithmetic will not stop until the function value of best individual less than a certain value or exceed the destined accelerate times. At this time, the currently group will be destined for the result of RAGA.

The above 8 steps make up of RAGA.

3.3 PPE model based on RAGA

Take projection function $Q(a)$ as the most target function in the PPE model and the projection $a(j)$ of each index as optimized variable. Through running 8 steps of RAGA, we can obtain the best projection direction $a^*(j)$ and projection value $z(i)$. Compare the $z(i)$ each other, we can obtain the evaluated result. Then, through comparing the distance between $z(i)$ and $Z(i)$, the smallest distance between any two samples, then, the number i is the soil sample grade.

4. Application example

Now, we use the data of literature [5] (Lu, 1999) to build up the RAGA-PPE model of water quality evaluation. The evaluated standards of the nutrition degree in lake see also to Table 1.

Table 1. Evaluation standard of lake nutrient states

Grade	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ	Ⅵ	Ⅶ	Ⅷ
Nutrition style	**Poor nutrition**	**Poor –medium nutrition**	**Medium nutrition**	**Medium-rich nutrition**	**Rich nutrition**	**Rather rich nutrition**	**Quite rich nutrition**	**Abnormity rich nutrition**
Chemistry oxygen demand（mg/L）	0.48	0.96	1.80	3.60	7.10	14.0	27.0	54.0
Total nitrogen（mg/L）	0.079	0.16	0.31	0.65	1.20	2.30	4.60	9.10
Total phosphor（mg/L）	0.0046	0.01	0.023	0.05	0.11	0.25	0.56	1.23

We can build up the PPE model of the evaluation standard about the nutrition states in lake based on MATLAB 5.3.

During the course of RAGA, the parent generation scale is 400 ($n=400$). The crossover probability is 0.80 (p_c =0.80). The mutation probability is 0.80 (p_m =0.80). The number of excellence individual is 20. α =0.05. Through accelerating 9 times, we can obtain the best projection value. That is 0.4203. The best projection direction: a^* =（0.5634，0.5553，0.6117）. Putting a^* into formula (2), we can obtain the projection value of each villages and towns. That are $z^*(j)$=（0.0121，0.0248，0.0491，0.1021，0.2020，0.4107，0.8409，1.7304）(Figure1, Table 2).

The PPC model based on RAGA of water quality is as follows.

$$y^*(i)=1.4138\ln z^*(i)+7.243 \qquad R^2=1.0000$$

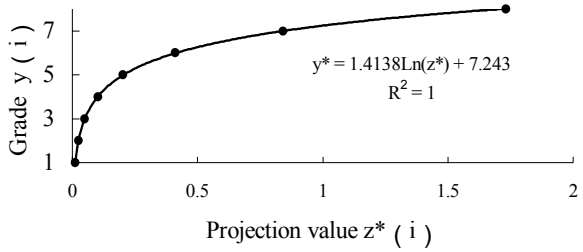

Figure1. The relation between water quality projection value and grade.

Table 2. Analyzing the error of PP model.

Experience value	1	2	3	4	5	6	7	8
Calculated value	1.0017	2.0163	2.9820	4.0170	4.9816	5.9849	6.9980	8.0183
Absolute error	0.0017	0.0163	-0.018	0.017	-0.0184	-0.0151	-0.0020	0.0183
Relative error	0.17%	0.815%	-0.60%	0.425%	-0.368%	-0.252%	0.029%	0.229%

The average absolute error is 0.0134, and the average relative error is 0.361%. We can see that the fit precision of RAGA-PPC model is rather high. So the PPE model can describe the relation between the evaluated index and water grade.

Now, we calculate the projection value of water quality about South-lake in Changchun during the course of June 1997 to May 1998 (Table 3).

Table 3. Evaluation results of nutrient states of South Lake water in Changchun.

Time	Chemistry oxygen demand（COD）（mg/L）	Total nitrogen（TN）（mg/L）	Total phosphor（TP）（mg/L）	BP Model[5]	Projection value Z*	Calculated value y*	RAGA—PPE Evaluated grade
1997.6	39.63	6.38	0.25	Ⅷ	1.0131	7.2614	Ⅶ
1997.7	40.73	2.85	0.36	Ⅷ	0.8682	7.0431	Ⅶ
1997.8	23.15	9.56	0.16	Ⅶ	0.9363	7.1499	Ⅶ
1997.9	35.24	5.76	0.13	Ⅷ	0.868	7.0429	Ⅶ
1997.10	32.42	2.40	1.15	Ⅷ	1.0531	7.3161	Ⅶ
1997.11	17.35	2.08	0.52	Ⅵ	0.5706	6.4498	Ⅵ
1997.12	40.67	2.05	0.63	Ⅷ	0.9317	7.1430	Ⅶ
1998.1	41.36	2.92	0.03	Ⅷ	0.7453	6.8273	Ⅶ
1998.2	40.19	4.50	1.49	Ⅷ	1.4205	7.7393	Ⅷ
1998.3	25.72	3.16	1.30	Ⅷ	1.0676	7.3417	Ⅶ
1998.4	33.32	2.57	1.01	Ⅷ	1.0178	7.2679	Ⅶ
1998.5	36.89	4.88	0.37	Ⅷ	0.9379	7.1524	Ⅶ
Year Average	32.35	4.83	0.62	Ⅷ	0.9758	7.2084	Ⅶ

From Table 3 we can know that the water quality in November, 1997 belongs to grade Ⅵ, and in February, 1998 belongs to grade Ⅷ during 1996/6 to 1998/5. The other months belong to grade Ⅶ. The

water quality in the whole year belong to grade Ⅶ. These mean the South-lake is in the states of graveness rich nutrition. The varied rule is rather accord with the practical condition. The PPE model has the same varied trend as the BP mode applied in literature (Lu, 1999). The time of wave crest and trough are the same. The results of other 10 months are litter higher than PPE model based on RAGA. The reason is obviously. When we are training the evaluated standard with BP network, there are artificial disturbance and subjectivity of determining the expected grade value output by BP network. Furthermore, because the distinguished ability of the BP network is rather low, so the result isn't accord with the RAGA-PPE model. But both of the two models can reflect the practical condition.

Furthermore, the best projection direction can reflect the influential degree of water quality grade caused by every evaluated index. The projection value is bigger, and it will have much influence for water quality evaluation. Thereby, we can verify the reasonability of the water quality evaluated standard. In the example, the best projection direction is a^* = （0.5634，0.5553，0.6117）. That means the three evaluated indexes have the same function in evaluating the water quality.

5. Conclusion

(1) Through applying PPC model, the author builds up the PPE model of evaluating the water quality. Several evaluation indexes have been taken as multi-dimension projection parameters to seeking the best projection direction. The best projection index function value can reflect the quality of each soil sample good or bad. Thus, we can avoid the disturbance by artificial factor to endow weight. The result is good.

(2) The author improves on SGA, and put forward a new method named RAGA through reducing the interval of excellence individual to accomplish the accelerate process. Thus, the method of RAGA can realize quick convergence and seeking the best result in the whole scope.

(3) Combing RAGA with PPE model, through using RAGA to optimizing the many parameters in the PPE model, we can obtain the best projection direction of evaluation index of each sample. Thus, the process of PPE modeling has been predigested. And the PPE model can be used in many other fields.

(4) The author put PPE model based on RAGA into the region of water quality evaluation. The PPE model can reflect the corresponding relation of non-linearity between classification number and projection value. The grade has been divided in focus. The precision of model is high. Furthermore, the best projection value can reflect the influential degree of each index during the course of total evaluation. Thus, the author provide a new method and thought for researching country energy programming.

Acknowledge

The financial support is provided by Chinese National "863" High-Technique Programme (No. 2002AA2Z4251-09); China Postdoctoral Science Foundation (No. 2004035167) ; Heilongjiang Province Youth Foundation (No. QC04C28); Natural Science Fund in China. (No. 30400275)

Correspondence to:

Qiang Fu, Wei Zu
School of Water Conservancy & Civil Engineering
Northeast Agriculture University
Harbin, Heilongjiang 150030, China
Telephone: 01186-451-55190298 (O)
Cellular phone: 01186-13936246215
E-mail: fuqiang100@371.net

References

[1] Jin Ju Liang, Ding Jing, 2000. Genetic arithmetic and its application to water science [M]. Chengdu: Si Chuan University publishing company: 42-7.

[2] Lu Wen Xi, Zhu Yan Cheng. Applying ANN to evaluate the nutrition states of south-lake in Changchun [J]. Geography Science 1999;19(5):462-5

[3] Mordern times mathematics notebook (volume of computer mathematics) [M]. Wuhan: Middle China science and technology publishing company: 2001:682-91.

[4] Zhang Xin Li. Projection pursuit ant its application to water resources [D]. Chengdu: 2000:67-73.

[5] Zhou Ming, Sun Shu Dong. The theory of genetic arithmetic and its application [M]. Beijing: National defense industry publishing company: 2000:4-7, 37-8.