Design of an Adaptive Gene Expression Programming Algorithms Based on Data Analysis

Wang Qin

Yang-En University, Quanzhou, Fujian, 362014, China

Keywords: Gene Expression Programming; Data Analysis; Convergence

Abstract: Gene expression programming is a new adaptive evolutionary algorithm based on the structure and function of biological genes. This algorithm has the shortcomings of slow convergence speed, easy to fall into local optimum and low fitting degree when solving specific problems. Based on data analysis, this paper proposes an adaptive gene expression program algorithm, which can adaptively adjust the crossover and mutation probability of the algorithm, thus effectively avoiding the sensitivity of artificial setting of initial parameters. It applies differential mutation search, chaotic reorganization and mutation operation, catastrophe operator to GEP. The results show that the algorithm not only improves the accuracy and convergence speed of the algorithm, but also effectively overcomes the immature convergence. The theory proves the global convergence of the algorithm. The improved genetic expression program has good performance.

1. Introduction

Gene expression programming is a new genetic algorithm based on genotype group and phenotype group proposed by Candida Ferreira, a Portuguese scientist. Nowadays, this method has been applied in many fields, such as function parameter optimization, evolutionary modeling, neural network, classification and TSP problems[1]. Unlike genetic programming, in gene expression programming, individuals use linear strings with fixed length to code and are represented as non-linear entities with different sizes and shapes. The algorithm has been successfully applied in many fields[2].

Zhou et al. showed that GEP can mine more streamlined and effective classification rules; Lopes and Weinert studied the application of GEP in symbol regression, and proposed a new system for analyzing symbol regression problems: EGIPSYS; Zuo et al. Using GEP for time series prediction, two prediction methods of GEP-SWPM and GEP-DEPM are proposed[3]. The experimental results show that the two methods have achieved good results in the prediction of sunspots. Huang Xiaodong et al. proposed a GEP-based method[4]. The function relationship discovery method - MEM method, that is, the domain domain expression mining. This method can deal with the relationship of consistent expressions and complex function relations with different domain expressions[5], and demonstrates that it has logarithmic order complexity; Wang Rui et al. used GEP to implement polynomial function decomposition and proposed the GPF method[6].

In the GEP algorithm, the parents' genes selected according to the fitness function are very close, so the resulting offspring are inevitably close to each other, and the degree of improvement expected is small. Thus, the unity of the gene model not only slows down the evolutionary process. And may lead to evolutionary stagnation, premature convergence to local best, resulting in low algorithm search performance[7]. Based on data analysis, this paper proposes an adaptive gene expression program algorithm, which can adaptively adjust the hybridization and mutation probability of the algorithm, thus effectively avoiding the sensitivity of artificially setting initial parameters. It applies differential mutation search, chaotic recombination and mutation operations, and catastrophic operators to GEP[8]. The research results show that the proposed algorithm not only improves the accuracy and convergence speed of the algorithm, but also effectively overcomes the immature convergence. The theory proves that the algorithm is globally convergent; the improved gene expression programming performance is good.

  DOI: 10.25236/mmmce.2019.035

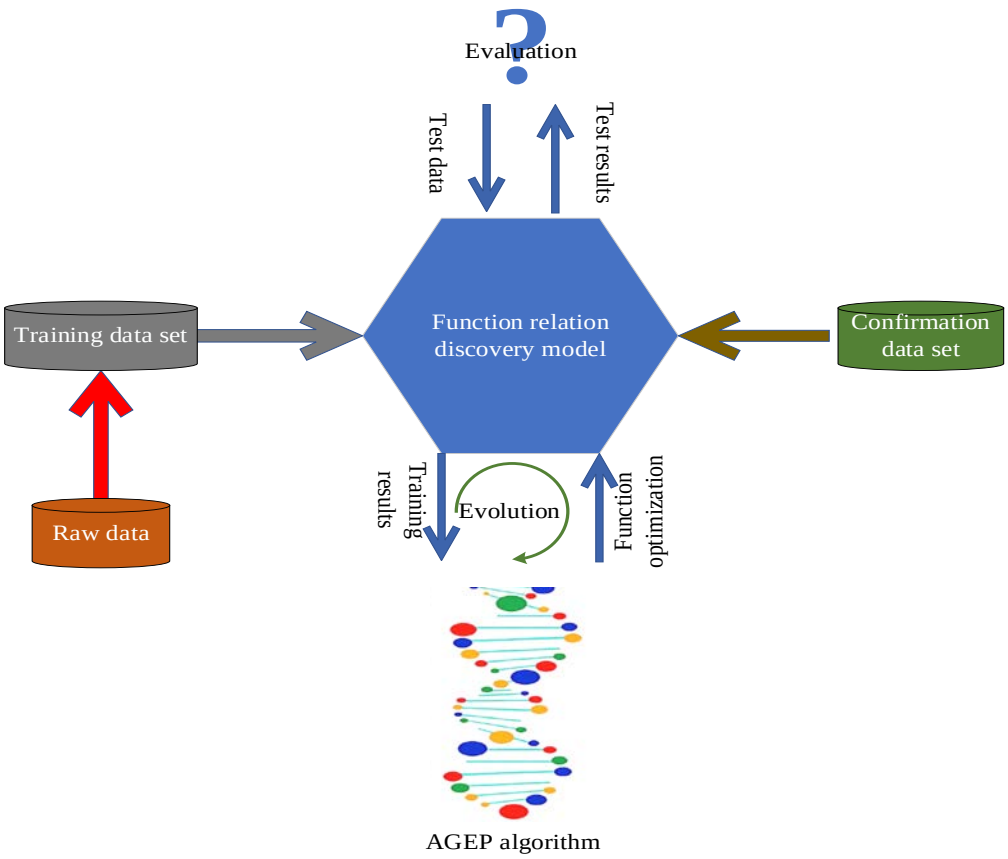

Figure.1 AEGP Functional Relation Model

2. Algorithmic design of adaptive gene expression program

2.1 Search technology based on differential catastrophe

Differential evolution is a joint evolution of Strom and Price . It is a fast evolutionary algorithm. It uses real-number coding and uses differential information between individuals to guide the search to generate new individuals[9]. For the differential evolution algorithm, in order to distinguish its method, the symbol “DE/a/b/c” is used. Where: DE is the algorithm; a is the mutated vector (which can be a random or optimal vector); b is the number of difference vectors; c is the crossover scheme (binomial or exponential). The mutation strategy “DE/rand/1” is a classic strategy [7], generating the intermediate vector v_i as follows.

$$v_i = x_{r1} + F \cdot (x_{r2} - x_{r3}) \quad (1)$$

In equation (1), $r_1,r_2,r_3 \in \{1,...,NP\}$;$r_1 \neq r_2 \neq r_3 \neq i$;$x_{r2} - x_{r3}$ are differential vectors, which can adaptively adjust the search during evolution. Direction and search step size; F is the scaling factor. However, differential heuristic performance can be significantly improved if heuristically controls the search direction more. To this end, a new search technology is proposed. The way to get $d_{i,j}$ is as follows:

$$d_{i,j} = \begin{cases} sign(u_{i,j} - x_{i,j}), if\ f(u_i) \leqslant f(x_i) \\ sign(x_{i,j} - u_{i,j}), if\ f(u_i) > f(x_i) \\ rand\ \text{int} \in \{-1,0,1,\}, \ else \end{cases} \quad (2)$$

2.2 Parameter adaptive control

Srinivas and Patnaik proposed a technique for adaptively controlling parameters based on adaptive values in genetic algorithm. This method can effectively avoid premature genetic

algorithm [4]. In this paper, the parameter adaptive technique is used to control the hybridization probability and mutation probability of GEP algorithm. The probability of hybridization (p_c) is calculated as follows:

$$p_c = \begin{cases} \dfrac{f_{\max} - f'}{f_{\max} - \bar{f}}, & f' \geqslant \bar{f} \\ 1.0, & \text{otherwise} \end{cases} \quad (3)$$

Among them, it is the fitness value of the current group's optimal individual, the mean value of all individual fitness values of the current group, and the fitness value of the better individuals of the two individuals involved in the hybridization. Formula (3) shows that when the fitness of the better individuals in the two individuals involved in the hybridization is better than the average of all the individuals in the population, the probability of hybridization is used. At this time, a small probability of hybridization is adopted to avoid destroying the existing good building blocks[6]. In particular, at the time, it was indicated that the hybridization operation was not performed on the optimal individual; at the time, it was indicated that the two individuals involved in the hybridization at this time were very poor, so the hybridization was performed according to the full probability[3].

2.3 Algorithmic design

The algorithm is as follows:
Begin
Randomly initialize the population P(0) while initializing dij;
Calculating the fitness of individual xi in P(0);
t=0;
Repeat;
Perform differential mutation search according to formula (3) to generate intermediate individual vi, group Part of P(t);
Perform chaotic recombination and mutation operations to recombine P(t);
Perform IS and RIS transformation operations to recombine P(t);
Calculate the fitness of individuals in P(t);
Select to generate the next generation of the parent from P(t) according to the selection strategy P(t+ 1);
Perform disaster operations
t= t+1;
Until the stop condition is met;
Systematic prediction using the best individuals in P(t);
End

3. Experiment and results

3.1 Experimental parameter setting

For the application data of the improved AGEP in function optimization, VC++ and Matlab are used as experimental platforms. Enter a set of regression data with two input variables x and y and one output variable z. The algorithm models the known input and output data, and then uses the established model to calculate the predicted value and calculates the relative error between the true value and the predicted value[7]. The conventional GEP algorithm, the improved GEP algorithm and the AGEP algorithm of this paper are tested, and the results are compared and analyzed[5].

Set the evolution operator: evolutionary algebra is 1 000, population size is 50, gene number is 5, head length is 6, IS and RIS transformation rate is 0.1, single point recombination, 2 point recombination rate pr and mutation rate pm chaotic Probability; the set of functions is {+,−,*,/,sqrt,ln,sin,cos,tan};Fitness function: mean variance, the range of random numbers is [−10,

10].

3.2 Experiment

The better model obtained by AGEP to calculate the data is show in Table 1.

Table.1. Experimental parameters

z	y	x	z	y	x
65	57	10	59	49	7
72	62	14	54	50	8
49	45	5	68	63	12
58	44	9	55	48	8
56	49	8	73	60	12
73	52	11	62	58	9

Table 2 Comparison of experimental results among GEP and improved GEP and AGEP

z (true value)	Predictive value		
	GEP	Improved GEP	AGEP
68.000 000	67.071 853	68.538 725	67.856 575
76.000 000	75.856 243	76.052 278	76.042 235
51.000 000	49.997 102	50.828 195	50.487 109
56.000 000	54.175 156	55.158 924	55.876 175
57.000 000	58.325 439	76.281 297	57.263 145
77.000 000	75.412 623	58.078 142	57.138 458
58.000 000	57.014 725	53.287 316	57.856 745
55.000 000	56.789 432	57.846 587	55.248 574
67.000 000	67.231 756	68.336 298	66.541 187
53.000 000	56.758 735	55.287 123	53.634 104
71.000 000	68.879 164	68.879 956	71.645 482
64.000 000	64.005 512	65.235 402	63.724 234

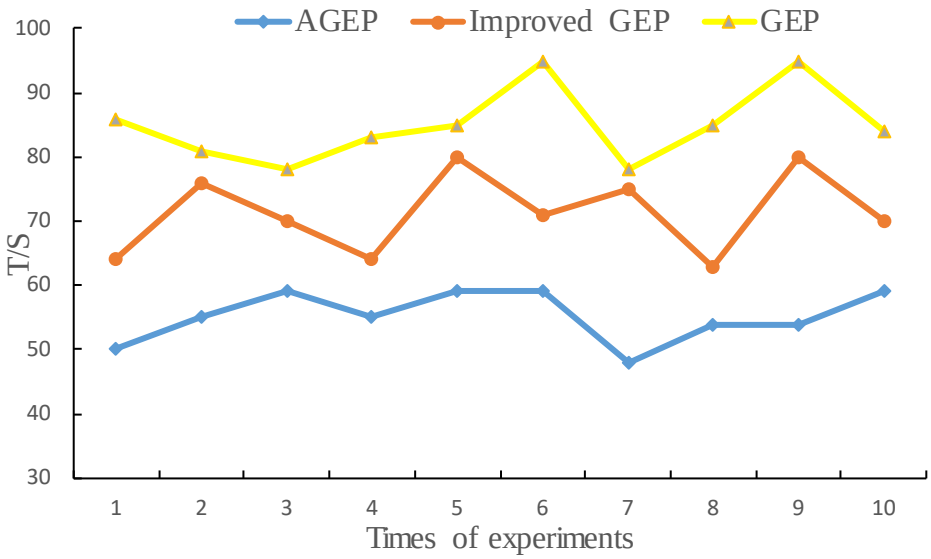

Figure.2 Comparison of time consumption among GEP and improved GEP and AGEP

It can be seen from Table 2 that GEP is better than GEP and the improved GEP algorithm. The average error of the AGEP algorithm is about 1/5 of the improved GEP algorithm in the literature, which is about 1/20 of the basic GEP algorithm. As can be seen from Figure 2: each time AGEP converges in 10 experiments.The speed is faster than the convergence of GEP and GEP in the literature, indicating that the AGEP algorithm has better convergence performance, robustness and

higher precision.

4. Conclusion

In this paper, an adaptive gene expression programming algorithm is proposed. Based on the original gene expression programming, this algorithm designs a differential mutation search, chaotic recombination, mutation operator and catastrophe operator suiTable for the characteristics of gene expression. This method can improve the convergence speed and accuracy of the algorithm, and effectively overcome the algorithm to fall into local optimum to achieve global optimization. However, the research of gene expression programming has just begun, far from systematic analysis methods and solid mathematical foundations like other evolutionary algorithms. Basic theoretical research has not yet made a breakthrough, and there is still a big gap between theory and application. It is believed that as scientific researchers continue to explore the depth, breadth and diversity of the theory, the theoretical basis of AGEP will be more solid and the application field will be more extensive.

References

[1] Qu L, Cheng H, Min Y. Adaptive Multi-phenotype Based Gene Expression Programming Algorithm [J]. Chinese Journal of Electronics, 2016, 25(5):807-816.

[2] Jie Y, Ma J. A hybrid gene expression programming algorithm based on orthogonal design[J]. International Journal of Computational Intelligence Systems, 2016, 9(4):778-787.

[3] Liu Z, Song Y Q, Xie C H, et al. Clustering gene expression data analysis using an improved EM algorithm based on multivariate elliptical contoured mixture models[J]. Optik - International Journal for Light and Electron Optics, 2014, 125(21):6388-6394.

[4] Lei Y, Li K, Zhang W, et al. Optimization of classification algorithm based on gene expression programming[J]. Journal of Ambient Intelligence & Humanized Computing, 2017(2):1-15.

[5] Zhou Yun, XU Jiucheng, XU Cunshuan. A cluster Algorithm for Time-course Gene Expression Profile Based on Affinity Propagation[J]. Journal of Henan Normal University, 2015, 46(7):1799-1809.

[6] Wang Z, Li G, Robinson R W, et al. UniBic: Sequential row-based biclustering algorithm for analysis of gene expression data:[J]. Sci Rep, 2016, 6(1):23466.

[7] Nepomuceno J A, Troncoso A, Nepomuceno-Chamorro I A. Integrating biological knowledge based on functional annotations for biclustering of gene expression data[J]. Computer Methods & Programs in Biomedicine, 2015, 119(3):163-180.

[8] Aghay S H, Kaboli, A, et al. Long-term electrical energy consumption formulating and forecasting via optimized gene expression programming[J]. Energy, 2017, 126:144-164.

[9] Wang M, Lu S. Validation of SuiTable Reference Genes for Quantitative Gene Expression Analysis inPanax ginseng[J]. Frontiers in Plant Science, 2015, 6(696).