A New Algorithm of Automatic Programming: GEGEP

Xin Du[1,2], Yueqiao Li[1], Datong Xie[1], and Lishan Kang[1,3]

[1] Department of Computer Science and Technology,
China University of Geosciences, Wuhan, China
`xindu79@126.com`
[2] Department of Information and engineering,
Shijiazhuang University of Economics, Shijiazhuang, China
`xindu79@126.com`
[3] State Key Laboratory of Software Engineering,
Wuhan University, Wuhan, China
`kang_whu@yahoo.com`

Abstract. Gene Expression Programming (GEP) has wide searching ability, simple representation, powerful genetic operators and the creation of high levels of complexity. However, it has some shortcomings, such as blind searching and when dealing with complex problems, its genotype under Karva notation does not allow hierarchical composition of the solution, which impairs the efficiency of the algorithm. So a new automatic programming method is proposed: Gene Estimated Gene Expression Programming(GEGEP) which combines the advantages of Estimation of Distribution Algorithm (EDA) and basic GEP. Compared with basic GEP, it mainly has the following characteristics: First, improve the gene expression structure, the head of gene is divided into a head and a body, which can be used to introduce learning mechanism. Second, the homeotic gene which is also composed of a head, a body and a tail is used which can increase its searching ability. Third, the idea of EDA is introduced, which can enhance its learning ability and accelerate convergence rate. The results of experiments show that GEGEP has better fitting and predicted precision, faster convergence speed than basic GEP and traditional GP.

Keywords: Gene Expression Programming; Genetic Programming; Gene Estimated Gene Expression Programming,; Estimation of Distribution Algorithm.

1 Introduction

Gene Expression Programming(GEP) was first proposed by Candida Ferreira [1]. In this new syste- m, the complex computer programs（the phenotype）evolved by GEP are totally encoded in simple strings of fixed length（the genotype）.

GEP is the inheritance and development of GA and GP, it synthesizes the merits of basic GA and traditional GP and has stronger ability of solving problems [2][3]. However, the learning procedure of GEP can be improved upon when dealing with complex problems with respect to both time efficiency and solution quality. The biological evolutionary process has revealed the principle of evolving from a self-contained functional

single cell to a well- developed entity with numerous specialized components [4]. We are naturally inspired to assume that solutions to complex problems might be built up incremen- tally from simpler elements. Although the phenotype of expression trees in GEP has retained the struct- ural representation from GP, the linear representation of the genotype conforms to Karva notation, under which the genotype-phenotype mapping mechanism does not guarantee that the levels of functional complexity in the phenotype are also directly reflected in the genotype. Since it is the genotype that is subject to the different genetic operations, it is difficult to follow the approach of incrementally forming solutions with the original GEP. Moreover, an evolved good functional structure is very likely destroyed in the subsequent generations not only by mutations but also by recombination and transpositions, which may require much additional computation to recover before an optimal solution is found. In view of this weakness, we propose a Gene Estimated Gene Expression Programming (GEGEP), it combines EDA with GEP . Compared with GEP, it mainly has the following characteristics: First, improve the gene expression structure, original gene structure is divided the head of chromosomes into a head and a body, which can be used to introduce the learning mechanism. Second, the homeotic gene which is also composed of a head, a body and a tail is used which can increase its searching ability. Third, the idea of EDA is introduced, which can enhance its learning ability and accelerate the convergence rate. We believe GEGEP benefits the evolution in terms of the convergence of a good functional structure.

The other parts of this article is organized as follows. Section2 explains GEGEP algorithm; Section3 proves the validity of the GEGEP algorithm through experiments; Section 4 presents some conclusions and ideas for future work.

2   Gene Estimated Gene Expression Programming

2.1   The Improvement of GEGEP

2.1.1   Improve the Structure of Gene Expression Programming

In GEGEP, gene head in GEP is divided into a head and a body. The head contains only functions, the body like the head of GEP contains symbols representing both functions and terminals, whereas the length of the tail is a function of head ,body and the number of arguments of the function with more arg- uments n(also called maximum arity) and is evaluated by the equation: tail =(head+body)×(n-1) + 1.

2.1.2   Improve the Structure of Homeotic Gene

Homeotic genes have exactly the same kind of structure as conventional genes of GEGEP and are built using an identical process. They are also composed of a head,a body and a tail.The heads only contain linking functions.The bodys contain linking functions, a special class of terminals:genic terminals representing conventional gene and random numerical constants.The tails contain obviously random numerical constants and genic terminals.
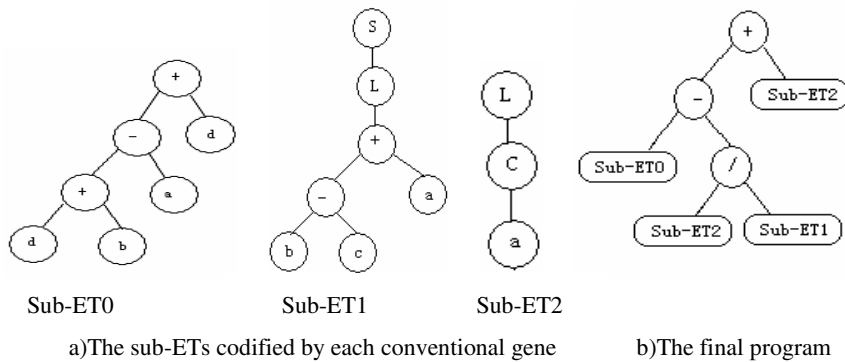
Sub-ET0   Sub-ET1   Sub-ET2

a)The sub-ETs codified by each conventional gene   b)The final program

Fig. 1. Expression of chromosomes

The head, body and tail of homeotic gene are respectively represented as homo_head, homo_body, homo_tail.The method of computation of homeotic gene is same as conventional genes. So the length of gene individual is evaluated by the equation:

length=(head+body+tail) *n_gene+ homo_head +homo_body + homo_tail, where n_gene represents the number of genes. For example, if we choose head=2, body=4,n=2,then tail = (head+body)*(2-1)+1=7,the length of conventional genes is 2+4+7 =13.In this case ,if n_gene=3, homo_head =1,homo_body=4,then the length of homeotic gene is equal to 11, the length of gene individual is equal to 50.

One such gene individual is shown below:

0123456789012 0 123456789012 0123456789012 01234567890

+-d+ adbc?d?c c SL+- abc?aa?bc LCa*b- abd? cca +-20 / 21?20?

Where ? represents random numerical constant, 0、 1、 2 respectively represent first、 second、 third conventional gene, L、 C and S respectively represent ln、 cos and sin function. It codes for three conventional genes and one homeotic gene. The conventional genes code as usual for three different sub-ETs. The homeotic gene controls the interactions between the different sub-ETs . The homeotic gene and conventional genes have same functional set and terminal set. The corresponding phenotypes are shown in Figure1.

2.1.3 EDA

EDA[5] is put forward from the viewpoint of statistics by MüHlenbein and Paaß, which is evolutionary algorithm based on probability analysis. EDA is similar to traditional genetic algorithm, the difference between them is that EDA selects evaluation and learning distribution to produce new offspring. New offspring are produced by Bayes network. The Bayes network is a directed acycline graph, which considers specific

interactions of the problem among the variables of chromosomes. It can make full use of realm knowledge and information of sampling datum with Bayes statistics. So it has strong learning ability and uses to model on discretional or continual polynomial in common.EDA contains three kinds of models: Univariate EDA Model, Bivariate EDA Model and Multivariate EDA Model which are shown in Figure2. Univariate EDA Model supposes that there have no interdependencies among variables. Bivariate EDA Model supposes that there have dependencies between pairs of variables. Multivariate EDA Model supposes that there have dependencies among variables greater than two.

For the convenience of computation, we only use Univariate EDA Model. The Model is shown in Figure 2(a). This basic thought of GEGEP is to carry on the statistics to the head and body department of each above operator of the conventional genes and homeotic gene , extracts its statistical probability, then carries on the mutation according to its statistical probability.
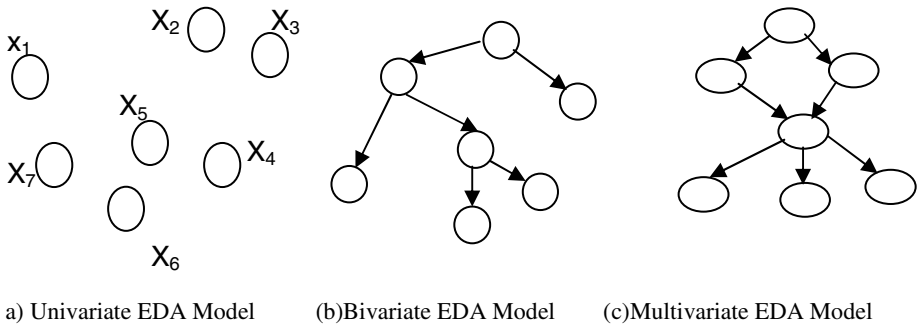

a) Univariate EDA Model   (b)Bivariate EDA Model   (c)Multivariate EDA Model

Fig. 2. Graphical representation of probability model among variables

The procedure of updating statistical information is as follows:

(1) Initialize fitness table

The size of fitness table is (n + 1) lines and L rows, where n expresses the size of function set, L expresses the total length of all he- ad and body added in the chromosome. For example: Function set = { +, -, *, /, Exp, Log, Sin, Cos, Sqrt }, the chromosome has 4 genes and 1 homeotic gene, the length of head of each gene is 3, the length of body is 7, then the size of total table is (9+1) *(5* (3+ 7)) = 10 *50. The initial value is ε ($\varepsilon = 0.1$) and initial frequency is once each generation.

(2) Initialize statistical information table

Statistical information table has the same size as fitness table, the initial value is also ε ($\varepsilon = 0.1$).The frequency is once every run.

(3) Renew fitness table

Renewing method is as follows: The fitness of each individual is added to the corresponding position of fitness table.

(4) Renew statistical information table

It directly adds the value of fitness table to original value of the corresponding position of statistical information table.

2.2   Fitness Function and Termination Condition

The fitness reflects whether the individual is good or bad. In this paper, the following fitness function is used:

$$f_i = \max\left(1000 - \sum_{j=1}^{n}\left(\left|\frac{C_{ij} - T_j}{T_j}\right| \cdot factor\right), 0\right)$$

Where n is the number of computation cases, Cij is the value returned by the individual program i for fitness case j and Tj is the target value for fitness case j. Factor is a proportional constant and is equal to 1000 dividing the number of records. Obviously, if the bigger fitness value is, then the better individual is.

The termination condition is the actual Evolutionary generation exceeding reserved evolutionary generation (Maxgenerations) or the best fitness having no change for L continuous generations：

If generation≥Maxgenerations1‖unchange_generation≥L

Then Terminates

The best solution or the approximate best solution is what we wanted when terminates.

2.3 Algorithm

PROCEDURE   GEGEP algorithm

begin

    Initialize P(0) ;

    Evaluate the fitness of P(0);

    Renew statistical information table;

    t:=0;

    repeat

      Pm(t):=mutation{P (t) } according to statistical information;

      Pc(t):=crossover{P (t) };

      Pt(t):=transpose{P (t) };

      P(t+ 1):=selection{P(t) };

      Renew statistical information table;

      t:= t+ 1;

    until termination condition ;

    output solution of Pbest;

end

3   Experiment

In order to justify the advantage of GEGEP as compared to traditional GP and basic GEP ,we first experimented on three algorithms with two function modeling . Later a suite of Symbolic Regression were conducted on the testing datasets .Last we experimented on another function modeling.

(1) Function modeling.

We select two cases from relevant references to check the validity of this method. Considering different impact of homeotic gene and conventional genes, the homeotic gene determines which gene is expressed in which cell and how they interact with one another. So the mutation rate is separated into homeotic gene rate and conventional generate.

The parameter setting of two experiments is same and is shown as Table1:

Table 1. Parameter Setting for GEGEP

Evolutionary generation	100000	One-point recombination rate	0.3
Population size	80	Two-point recombination rate	0.3
Mutation rate of homeotic gene	0.044	IS transposition rate	0.1
Mutation rate of conventional genes	0.044	RIS transposition rate	0.1
Gene recombination rate	0.1	Gene transposition rate	0.1
Conventional gene Numbers	4	Homeotic gene's head length	4
Conventional genes' head length	2	Homeotic gene's body length	20
Conventional genes' body length	15		

The calculation result of Y by GP and GEP are respectively obtained from [6] and [7]. There are six main factors: the depth of coalface x_1 (m) ,the height of coalface x_2 (m) , the amount of coalfaces x_3 (m3/t) , the layer interval between working coalface and neighboring coalface x_4 (m) , average day progress of working face x_5 (m /d) and average output per day x_6 (t/d) as the main input variates ,the amount of gas emitted from coalfaces(Y) as the output variate. So the termination set is defined as T = {x_1, x_2, x_3, x_4, x_5, x_6, ? },where ? represents random numerical constant and is chosen from the interval [-10, 10].The function set is defined as F = {+, −, *, ÷, sqrt, exp, sin, cos, ln}.The aim is to obtain a prediction function for Y with x_1, x_2, x_3, x_4, x_5 and x_6 being the parameters: Y =μ (x_1, x_2, x_3, x_4,

x 5, x 6). Fifteen cases from sample 1 to 15 in table2 are used as training datum to build the prediction model while three cases from sample 16 to 18 in table3 are used as test datum.

The function evolved by GEGEP is：

$$y=Ln(Exp((Sin((Sin(((Exp(x3)+x3)-x6))/2.3661))+x2)))+Exp((Sqrt(Sin(Sin(Ln(x4))))$$
$$+(Exp((Sin(Sin((x3*x6)))+Sin(x4)))/((x3-Ln (Sin(Sqrt(x5))))+ x4))))$$

Table 2. Comparison of fitting results by GP ,GEP and GEGEP

sample	x_1 (m)	x_2 (m)	x_3 (m³/t)	x_4 (m)	x_5 (m/d)	x_6 (t/d)	Calculation result of y （m³/min）				Relative error of Y		
							Y	GP	GEP	GEGEP	GP	GEP	GEGEP
1	408	2.0	1.92	20	4.42	1825	3.34	3.18	3.27	3.37	0.048	0.020	0.00885
2	411	2.0	2.15	22	4.16	1527	2.97	2.79	3.29	2.93	0.061	0.108	0.01420
3	420	1.8	2.14	19	4.13	1751	3.56	3.89	3.47	3.56	0.092	0.024	1.62E-05
4	432	2.3	2.58	17	4.67	2078	3.62	3.46	3.71	3.71	0.044	0.026	0.02398
5	456	2.2	2.40	20	4.51	2104	4.17	3.98	4.10	4.15	0.046	0.018	0.00579
6	516	2.8	3.22	12	3.45	2242	4.60	4.78	4.68	4.64	0.039	0.017	0.00913
7	527	2.5	2.80	11	3.28	1979	4.92	4.88	5.03	4.98	0.008	0.022	0.01178
8	531	2.9	3.35	13	3.68	2288	4.78	4.38	4.84	4.87	0.084	0.012	0.01833
9	550	2.9	3.61	14	4.02	2325	5.23	5.1	5.19	5.21	0.025	0.008	0.00291
10	563	3.0	3.68	12	3.53	2410	5.56	5.78	5.42	5.57	0.040	0.024	0.00102
11	590	5.9	4.21	18	2.85	3139	7.24	7.09	7.35	7.24	0.021	0.015	0.00056
12	604	6.2	4.03	16	2.64	3354	7.80	8.15	7.49	7.78	0.045	0.040	0.00291
13	607	6.1	4.34	17	2.77	3087	7.68	7.36	7.58	7.68	0.042	0.013	3.1E-05
14	634	6.5	4.80	15	2.92	3620	8.51	8.91	8.30	8.52	0.047	0.024	0.00105
15	640	6.3	4.67	15	2.75	3412	7.95	7.43	7.94	7.98	0.065	0.001	0.00428

Table 3. Comparison of predicted results by GP,GEP and GEGEP

sample	x_1 (m)	x_2 (m)	x_3 (m³/t)	x_4 (m)	x_5 (m/d)	x_6 (t/d)	Calculation result of y （m³/min）				Relative error of Y		
							Y	GP	GEP	GEGEP	GP	GEP	GEGEP
16	450	2.2	2.43	16	4.32	1996	4.06	4.32	4.11	4.01	0.064	0.012	0.01268
17	544	2.7	3.16	13	3.81	2207	4.92	5.15	4.93	4.91	0.046	0.002	0.00105
18	629	6.4	4.62	19	2.80	3456	8.04	7.79	8.01	7.80	0.031	0.004	0.03037

From Table2, we can see that the largest relative errors of GP, GEP and GEGEP are 9.2%, 10.8% and 2.398% respectively. Meanwhile, among the 15 training samples, the result of GEP has smaller relative errors than the one of GP on 13 samples, and the result of GEGEP has smaller relative errors than the one of GEP on 13 samples. It's obvious that concerning the goodness of fit, GEP outperforms GP and GEGEP outperforms GEP.

From table3, we can see that GEGEP has the higher prediction results and a smaller relative error of Y than the one of GP and GEP on three and one sample respectively.

(2) Function modeling.

The data comes from paper[8], which describes the GDP of China in some periods .

The termination set is defined as T = {K, L, ? }, where ? represents random numerical constant and is chosen from the interval [-10, 10].The function set is defined as F = {+, −, *, ÷, sqrt, exp, sin, cos, ln, tan}. The parameter setting is same as experiment1.

The aim is to obtain a prediction function for GDP with K and L being the parameters: GDP =μ (K, L). Fifteen cases from the year 1980 to 1994 in table4 are used as training datum to build the prediction model, while two cases from 1995 to 1996 in table5 are used as test datum.

The second column（K）in table3 and table4 describes the sum of net value of fixed assets and the average balance of floating assets in the corresponding year. The third column（L）describes the number of employed person, including all kind of workers and peasants. The fourth column（calculation result of GDP）describes the actual statistical data of GDP in the same year.From Table4, we can see that the largest relative errors of GP, GEP and GEGEP are 8.29%, 3.182% and 3.177% respectively. Meanwhile, among the 15 training samples, the result of GEP has smaller relative errors than the one of GP on 13 samples, and the result of GEGEP has smaller relative errors than the one of GEP on 6 samples. It's obvious that concerning the goodness of fit, GEP outperforms GP and GEGEP outperforms GEP.From table5, we can see that GEGEP has the higher prediction results and a smaller relative error of GDP than the one of GP and GEP on one sample.

Table 4. Comparison of fitting results by GP ,GEP and GEGEP

Year	K (109 yuan)	L (104)	Statistical data of GDP (109 yuan)	Calculation result of GDP（109 yuan)			Relative error of GDP		
				GP	GEP	GEGEP	GP	GEP	GEGEP
1980	461.67	394.79	103.52	100.56	103.521	103.800	0.0286	0.00001	0.00271
1981	176.32	413.02	107.69	111.29	107.696	107.504	0.0309	0.00006	0.00145
1982	499.13	420.50	114.10	120.02	114.687	115.421	0.0519	0.00514	0.01158
1983	527.22	435.60	123.40	133.63	123.374	123.460	0.0829	0.00021	0.00049
1984	561.02	447.50	147.47	147.95	142.777	147.238	0.0032	0.03182	0.00157
1985	632.11	455.90	175.71	171.68	173.509	175.821	0.0229	0.01253	0.00063
1986	710.51	466.94	194.67	199.85	193.524	193.802	0.0266	0.00589	0.00446
1987	780.12	470.93	220.00	222.45	217.766	217.706	0.0114	0.01015	0.01043
1988	895.66	465.15	259.64	252.39	260.763	256.438	0.0279	0.00433	0.01233
1989	988.65	469.79	283.34	282.81	287.707	286.217	0.0019	0.01541	0.01015
1990	1075.37	470.07	310.15	308.42	310.990	309.670	0.0056	0.00271	0.00155
1991	1184.58	479.67	342.75	349.05	346.464	347.097	0.0184	0.01084	0.01268
1992	1344.14	485.70	411.24	403.05	410.710	410.960	0.0199	0.00129	0.00068
1993	1688.02	503.10	536.10	529.49	531.169	553.134	0.0123	0.00920	0.03177
1994	2221.42	513.00	725.14	714.81	728.118	729.770	0.0142	0.00411	0.00639

The function evolved by GP is：

GDP=exp (ln (K)-576.032/L)-ln (K+L)

The function evolved by GEP is：

GDP=tan(sin(sin(L)/cos(K)))+sin(tan(L))*9.16+log(abs(tan(L)))+K*log(log(abs (sqrt(L-6.44)+L-K)))/6.32+log(14.46*sqrt(L)+sin(K))/sin(log(K+L))+tan(log(L+5.08) *L3/-9.58)

The function evolved by GPGEP is：

GDP=2*((tan(-2.95022)*K)-sqrt((2*K))-8.92026-2*(((cos(sqrt((((-0.36897)+K)))* (-8.17927))/(2.44667))-tan(ln(K-5.137215)))-(tan(0.450852-K)-0.99622)+tan(exp(cos (ln(K))))+exp(tan(K))/cos(cos(L))+tan(((cos(sqrt(K-0.36897))*(-8.17927))/(2.44667)) -tan(ln(K-5.137215)))

Table 5. Comparison of fitting results by GP ,GEP and GEGEP

Year	K (109 yuan)	L (104)	Statistical data of GDP (109 yuan)	Calculation result of GDP（109　yuan)			Relative error of GDP		
				GP	GEP	GEGEP	GP	GEP	GEGEP
1995	2843.48	515.30	920.11	921.64	922.012	921.342	0.0017	0.00207	0.00134
1996	3364.34	512.00	1102.10	1083.91	1107.010	1126.604	0.0165	0.00446	0.02223

4   Conclusions

This paper puts forward a new algorithm GEGEP, which takes GEP as foundation, improves gene structure and introduces the learning mechanism. The learning mechanism benefits preserving good structure which leads to a better evolutionary process. So it has the potential of identifying useful structural information emergent in the evolutionary process. Practical examples show that this alg- orithm possesses the advantages of automation of the modeling process, more flexible and various model structures, wider range of applications , faster speed of convergence and higher precision of fitting and predicted datum.

Future research on GEGEP will mainly focus on the following:

(1) Here we only consider Univariate EDA Model and do not consider Bivariate EDA Model and Multivariate EDA Model. So next we will try to use Bivariate EDA Model or Multivariate EDA Model. (2) The speed of algorithm has decreased because of large quantities of computation spending for statistics. We will adopt some methods to improve. (3)We will realize parallelism which can further improve the ability of solving complex problems. (4)Here we only realize one-layer call model , next we will realize multilayer call model.

Acknowledgments

References

1. Ferreira, C. Gene Expression Programming: a New Adaptive Algorithm for Solving Problems. Complex Systems, 13, 2(2001), 87-129.
2. Ferreira C. Gene expression programming[M] . Portugal, An- gra do Heroismo ,2002.
3. Ferreira C. Gene expression programming in problem solving [A]. 6th Online World Conference on Soft Computing in Indu- strial Applications[C] . 2001.
4. Xin Li, Chi Zhou, Weimin Xiao, Peter C. Nelson. Prefix Gene Expression Programming. Genetic and Evolutionary Computa- tion Conference (GECCO)'05, Washington, DC, USA., June 25-29, 2005,
5. P. Larra˜naga and J.A. Lozano. Estimation of distribution alg- orithms. A new tool for evolutionary computation. Kluwer Ac- ademic Publishers, 2001.
6. Zhao CY, Yuan XG，Sun JB. Application of Genetic Progr- amming to Predicting the Amount of Gas Emitted from Coal Face [J]. Journal of Basic Science and Engineering. 1999,7(4): 387-392.
7. Li Q, Cai ZH，Zhu L, Zhao SY. Application of Gene Expr- ession Programming in Predicting the Amount of Gas Emitted from Coal Face[J].Journal of Basic Science and Engineering. 2004.3(12).49-54.
8. CAI Zhihua, JIANG Siwei, Zhu Li , GUO Yuanyuan . A Novel Algorithm of Gene Expression Programming Based on Simul-ated Annealing. International Symposium on Intelligence C-omputation & Applications [C], 2005.605-610