Applied Mathematics & Information Sciences
An International Journal

http://dx.doi.org/10.12785/amis/090427

A heavy Metals Morphology Prediction Modelling Method based on JM-GEP

Yongqiang ZHANG[1,*] *and Junxia LI*[2]

The information and electricity-engineering institute, Hebei University of Engineering, Handan, 056038, P. R. China

Received: 28 Oct. 2014, Revised: 28 Jan. 2015, Accepted: 29 Jan. 2015
Published online: 1 Jul. 2015

Abstract: In this paper, an improved GEP (Gene Expression Programming based on Jumping Genes, JM-GEP) is proposed in consideration of the morphology of heavy metals (HM) changed over time, on which a new heavy metal prediction method has been put forward. Jumping operator is the key point to JM-GEP, in which the jumping operators use self-adaptive jumping probability to keep population diversity and study the convergence property of the optimal retention strategy. Aiming at the improved GEP, we raised a heavy metals modelling method based on JM-GEP. The simulation results show that the new model, compared with traditional methods, has excellent goodness of fit to HMFT characteristic function, and find out its global optimal. The new method proved to be widely used for researching other time sequences problems.

Keywords: HMFT, Gene expression programming, JM-GEP, Heavy metal prediction model

1 Introduction

At present, the disposal process of municipal solid waste (MSW) is incineration. As the result of the incineration process, different solid residues, such as bottom ash, fly ash, and air pollution control residues are generated. Likewise, it could be of significance in evaluating the possibility of reusing bottom ash as a secondary material, taking into account the potential environmental impact. However, bottom ash has a high content of potentially harmful heavy metals relative to soil, environment and people. Then, proposing a precise heavy metals prediction model to predict HM mechanical properties is becoming more and more important. In the early 1980s, many models of heavy metals form have been put into use. Currently, the research on heavy metals prediction model has the following three categories: direct determination method [1,2], model calculation method[3,4,5,6], chemical extraction method (continuous extraction)[7,8,9]. While paper [1]is to bind constants of lead by humic and fulvic acids studied by Anodic Stripping Square Wave Voltammetry; paper[2] is to bind concentrations and ionic strength on copper lead by humic and fulvic acid; Paper[3,4,5,6] is to predict the characteristics of HM by Chemical Equilibrium Calculation model; Paper[7,8,9] is to survey the concentration and PH decreasing with time(after 12 weeks)(Lead,Zinc,Copper) to predict HM model. However, due to lots of factors affecting heavy metals form(PH ORP organic/inorganic absorption, etc.), this paper is going to analyse the mechanisms of heavy metals migration and transformation. As the result of the three studies is not uniqueness and completeness, it brought an unprecedented difficulty for the study. This article put forward a new heavy metals prediction modelling method based on gene expression programming (GEP) by the form of heavy metals in municipal solid waste incineration (MSWI) bottom ash and the morphological transformation of time series data mining. In order to better predict the characteristics of the morphological transformation with time, this article first put the jumping gene into GEP so that it has higher and faster searching ability. So an improved method based on jumping gene expression programming (JM - GEP) predictive model of heavy metals is proposed. This improved algorithm is analysed through the following four aspects: JM-GEP population diversity analysis; the complexity of the algorithm and convergence; JM - GEP on morphological transformation with time series data mining model parameters; comparing with traditional GEP model and other models to evaluate the new model.

* Corresponding author e-mail: yqzhang@hebeu.edu.cn

2 GEP foundation

GEP invented by Ferreira [10] in 1999, is an inevitable trend in the development of GAs and GP. The core of GAs randomly generated a simple linear binary chromosome string; GP is a tree structure computer program on the basis of the GAs.

While GEP is expression trees a tree structure computer program which is on the basic combination of Gas and GP. The core algorithm of GEP mainly includes fitness function, selection operator, mutation operator, crossover and transposon operators, etc. Though each operator is essential to maintain the population diversity, there are still many problems to process carefully, such as poor capability of global optimization, local convergence and premature easily. In order to maintain GEP population diversity and avoid falling into local convergence, a lots of improved evolutionary algorithm have been came out to adjust evolutionary parameters, like mutation rates[11,12,13,14] and selection pressure[15]and so on. Here jumping genes enter gene expression programming firstly used to propose newly JM-GEP.JM-GEP algorithm is a newly adduced self-adaptive strategies, which is applied to dynamically adjust the parameters of genetic algorithms for the purpose of enhancing the performance, maintaining JM-GEP population diversity, avoiding local convergence and prematurity.

2.1 Fitness function selection

The design approach fitness function is the key to success to solve the problem. In GEP, there are several approaches can be used to evaluate how good these models are. Some of them are based on the absolute error between predicted and target values. Others are based on relative error rather than absolute error. The following three ways are usually adopted [16].

$$f_i = \sum_{i=1}^{n} (R - |P_{(i,j)} - T_j|) = 0 \quad (1)$$

$$f_i = \sum_{i=1}^{n} (R - \left| \frac{P_{(i,j)} - T_j}{T_j} \times 100 \right|) = 0 \quad (2)$$

$$if \quad N \geq \frac{1}{2}n, then \quad f_i = N, else \quad f_i = 1 \quad (3)$$

Where R is the selection range, $P_{(ij)}$is the value predicted by the individual program i for fitness case $j$out of n fitness cases and T_j is the target value for fitness case j. And$N$ is the number of correct cases. Noting that equation(11) and (2) can be used to solve any symbolic regression problem, while (3) to logic problems. In the design of fitness function, the goal is very clear to make the evolutionary direction of the system in accordance with requirements.

2.2 Transposition

Transposition operator[19] is specific operator among all genetic operators. Transposition operator acts on a single or double chromosome. Three transpositions in GEP shown in this: Starting position is sequence elements of the function or terminal symbols insert the outside of the root node in the head (Insertion Sequence Elements). Starting position is sequence elements of the function or terminal symbols insert the root in the head(Root Insertion Sequence Elements). Gene inserts the starting position.

Figure1 illustrates the parent p produces children s through IS element (IS Transposition). It has chosen IS length is 3 $(9^{th} \sim 11^{th}, bold)$, then insert into 3^{th} position and $5^{th} \sim 7^{th}$ codes are truncated.

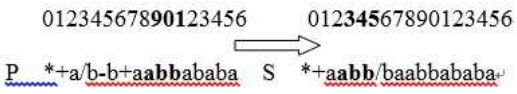

Fig. 1: IS Transposition

Here we define RIS length is $3(3^{th} \sim 5^{th} position, bold)$. Figure2 illustrates the parent chromosome p produces children s through RIS(RIS Transposition). And the five to seven codes are truncated.

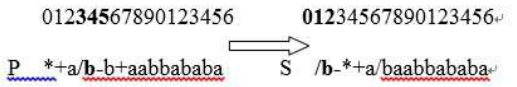

Fig. 2: RIS Transposition

Where in Gene Transposition, the whole gene act as a transposon insert into starting position, then other gene backward in turn. The difference is transposon was deleted to ensure chromosome length does not changed. But Gene Transposition only acts on multiple genetic chromosomes. Figure3 illustrates the Gene Transposition Operator.

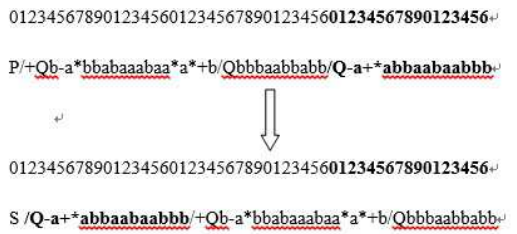

Fig. 3: Gene Transposition

2.3 Jumping operator

Jumping genes, just as its name, can be able to replicate itself and jump from one place to another. But in fact, they are some special DNA fragments on chromosome, which is not only can ”jump” from a place to another on the same chromosome, but also can jump between the chromosomes. This jumping can lead to some changes on chromosome, such as genetic mutation, number of genes alter. But in order to maintain the effectiveness in the population, here we do not change the chromosome length. In the early 1940s, American McClintock, put forward the jumping genes. This view broken the concept of a stable chromosome. Although it was not recognized, it was more common than think acturally and almost all person has a unique ”jumping genes”.

Jump operator has worked on a single or even doubled chromosomes. There exist two ways of jumping, copy and paste operation, which is to copy transposon gene on a chromosome, and insert other chromosomes in this article, and the inserting position of the original gene segment are covered[16], cut and paste operations, which is to cut transposon genes on a chromosome, and paste it into the insert position or other chromosomes where the location of cutting can be replaced by subsequent genes. At the same time, cutting gene segment length is equal to the former on the other chromosome and paste it into the chromosome of insert position. We know the jump location and insert position are randomly selected. But the number and the length of the jumping genes is artificial. Jumping behavior is as shown in figures 4,5.

Genetic operators play an important role in gene expression programming. But once the designing is unreasonable, it is easy to reduce the population diversity, relapsing into the local convergence and premature at the end of evolution. So this article put jumping genes into GEP, and jumping operate is founded behind selection operator and mutation operator so that these defects can decrease as low as possible. The basic steps as shown below. Step 1: The program will generated N populations randomly $C_1, C_2, \ldots C_N$. Then evaluate the fitness function of every generation, then we rank them from the highest fitness function to lowest, a group of 20, calculate the fitness function in each group. Step 2: select individuals with high fitness in each group to carry out mutation operation. Step 3: Suppose jumping rate $P_{jm} = 0.1$, the choice of ways of jumping rate was $JmPnStprand$. If the rate of individuals,P_1 is lower than P_{jm}. Jumping behaviour is coming, and then cut and paste operation occurred when $JmPnStprand$ is greater than P_{jm}. Otherwise, copy and paste operation occurred. Here jumping genes and the choice of location as well as the length of the insert position are randomly generated.

3 JM-GEP algorithm description

In this paper, we proposed an evolutionary algorithm JM-GEP to maintain the population diversity, and prevent or reduce premature and jump out of local optimal solution through researching the GEP with reference to the thought of Chen[17]. The validity of the algorithm JM-GEP is proved through population diversity analysis and complexity analysis, convergence analysis and so on. A detailed description as Figures 6,7,8,9.

(1) Fitness Function

The two evaluation models based on error proposed by Candida has their own inherent shortcomings [18]. In statistics, R^2(Coefficient of Determination) is used to evaluate the excellent or bad result of using HM prediction model to forecast the form of HM, lie on the adjacent degree of two sets of data. The calculation formula is as below.

$$R^2 = \frac{\sum_{i=1}^{n}(f_i' - f_i)^2}{\sum_{i=1}^{n}(f_i - f_{ave})^2} \tag{4}$$

Where f_i is the real observed value, and f_i' is the regressed value, f_{ave} is the average one of observed values.

(2)Jumping Genes operator

As we all know from here, the rate of jumping, the choice of the jumping location and insert location are all selected randomly so that it greatly reduces the jumping probability as well as population diversity. So we set dynamic jumping probability in this paper, in order to make jumping genes operator self-adaptive. First, we rank them from the highest to lowest according to the fitness function. Then every 20 as a group and each group of jumping genes probability function is designed as follows.

$$p_i = \alpha\left(1 - \frac{i}{N/20}\right) \quad i = 1,2,\ldots N/20 \tag{5}$$

Where α is a constant before evolutionary with a range of $(0, 0.15)$.

3.1 JM-GEP Population Diversity Analysis

The best measurement population diversity can be represented by Shannon entropy function of generated uniformity. **Definition 1**[19] For an uncertain system, suppose a random variable represents its state characteristics , a value of random discrete variable is $X = \{x_1, x_2, \ldots x_n\}, n \geq 1$, which the probability value is .So the average entropy is $H(X)$ that represents the uncertainty of program. The following is showing $H(X)$.

$$H(X) = -\sum_{i=1}^{n} p_i log_a p_i \tag{6}$$

Where $p_i \in R, p_i(= p(x_i))$is the probability of information symbol x_i appeared[20,21].When $p_i = 0$,

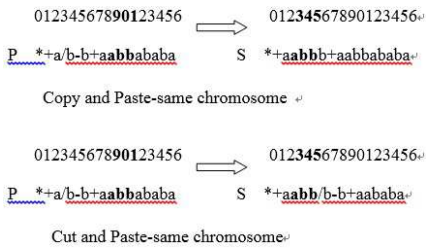

Fig. 4: Same Chromosome Jumping Operator

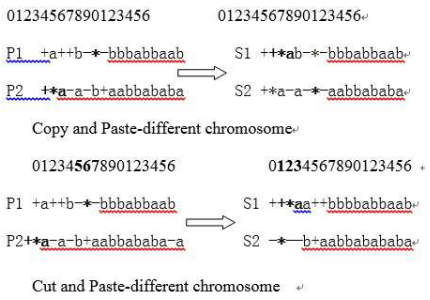

Fig. 5: Different Chromosome Jumping Operator

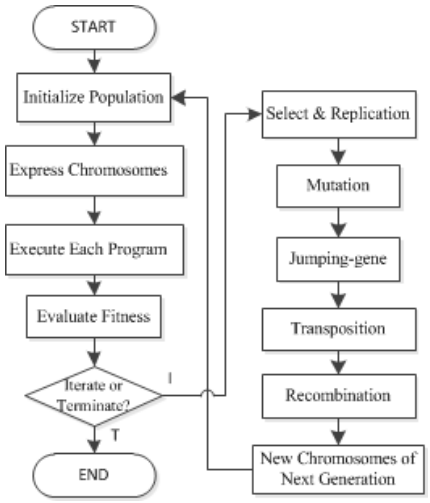

Fig. 6: JM-GEP algorithm

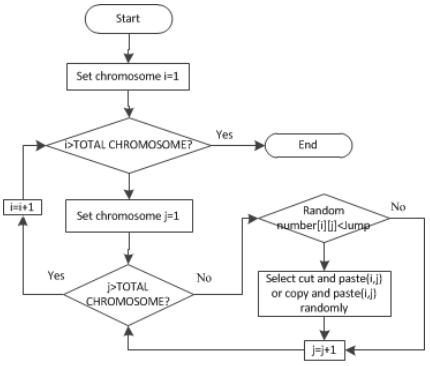

Fig. 7: Jumping genes detailed flow diagram

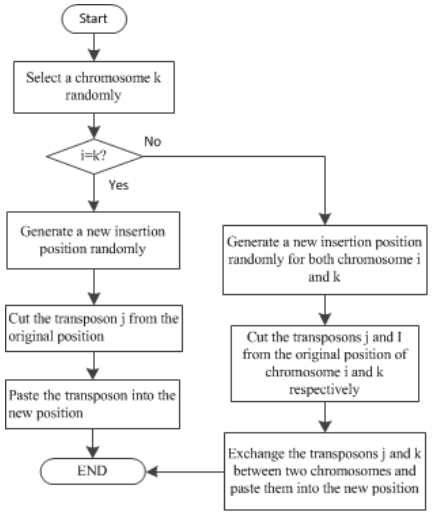

Fig. 8: cut-paste operation

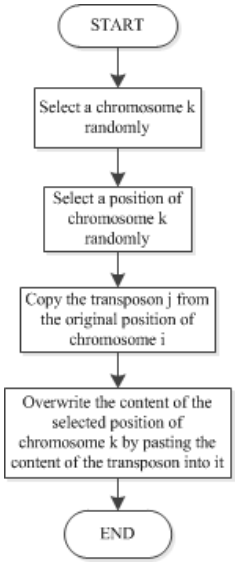

Fig. 9: copy-paste operation

Table 1: The data series of HM form changed over time

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
t_i	0	4	8	18	24	32	36	44	48	52	56	60	64	68	72	76
T_i	0.01	0.05	0.08	0.07	0.09	0.05	0.03	0.01	0.09	0.08	0.06	0.11	0.11	0.05	0.03	0.01

then $p_i log_a p_i = 0$. From Formula (6), we can know, Shannon entropy is only related to sate probability vector, $\{p_1, p_2, \ldots p_n\}$, and not related to its concreteness value. In this paper, whatever the jumping rate is change among signs in same gene sequence, the sequence is always balancing [22]. The equation is as follows.

$$p_i = \frac{1}{F+T} \quad (i = 1,2,\ldots F+T)$$

Where, F is the number of function set, T is the number of terminal set. So according to the maximum entropy principle, the entropy is not only maximum value, but also not changes with time. The formula (7) is as shown below.

$$H = -\sum_{i=1}^{F+T} p_i ln p_i = -\frac{1}{F+T} ln \frac{1}{F+T} = ln(F+T) \quad (7)$$

In this paper, L(the gene sequence length of HM changes with time) is a finite value. And the frequency of each sign appeared in gene sequence may throw off $1/(F+T)$ so that the comentropy will respond to jumping rate and changes with copy. Supposed a gene sequence, the length of it is a constant. Likewise, the number of each symbol appeared in the gene sequence is $f_{i,t}(i = 1,2,\ldots n) \quad (n = F+T)$. And entropy is H_t; So the gene sequence entropy is as follows when t equals 0.

$$H_0 = -\sum_{i=1}^{n} \frac{f_{i,0}}{L} ln \frac{f_{i,0}}{L} \quad (8)$$

Suppose jumping rate is p_{jm}, at the$(t+1)^{th}$, is the number of symbols appeared in sequence.

$$f_{i,t+1} = n f_{i,t} - \sum_{ji} f_{j,t} p_{jm} \quad (9)$$

Because $f_{i,t} + \sum_{j \neq i} f_{i,t} = \sum_{i=1}^{n} f_{i,t} = L, t \in [0, +\infty]$, and then formula 9 above can be formula 10).

$$\triangle f_{i,t} = n f_{i,t} - L \quad (10)$$

Due to the time of every evolution generation is smaller than the whole evolutionary time. And we image time is continuous, so the equation(9)can be equation (11).

$$\frac{d f_{i,t}}{dt} = n f_{i,t} - L \quad (11)$$

From the equation (5) above shows that jumping rate is only relevant to the individuals and parent chromosomes, and has no relation with time. Therefore, jumping rate can be regarded as constants. We know that the number of the i^{th} symbol appeared in the sequence is $f_{i,0}$when t is equal to0, you can get the equation (12) by solving differential equation (11).

$$f_{i,t} = \frac{L}{n + f_{i,0} - \frac{L}{n}} e^{-nt} \quad (12)$$

Where $t$tends to be infinity, $f_{i,t}$tends to be $\frac{L}{n}$. So whenever t was, gene sequence entropy was always formula (13).

$$H_t = -\sum_{i=1}^{n} \frac{\frac{L}{n} + \left(f_{i,0} - \frac{L}{n}\right)^{e^{-nt}}}{L} ln \frac{\frac{L}{n} + (f_{i,0} - fracLn)^{e^{-nt}}}{L} \quad (13)$$

Population diversity can then analyze the sift in gene sequence entropy H_t.

Suppose $F_{i,t} = -\frac{f_{i,t}}{L} ln \frac{f_{i,t}}{L}$then $\frac{dF_{i,t}}{dt} = \frac{n f_{i,t} - L}{L} ln \frac{L}{e f_{i,t}}$, so

$$\frac{dH_t}{dt} = \sum \frac{n f_{i,t} - L}{L} ln \frac{L}{e f_{i,t}} = ln \prod_{i=1}^{n} \left(\frac{L}{e f_{i,t}}\right)^{\frac{n f_{i,t} - L}{L}} \quad (14)$$

$f_{i,t} \leq L \ \frac{f_{i,t}}{L} \leq 1 \ \frac{L}{e f_{i,t}} \geq \frac{1}{e}$ In the formula (14), when $L - n f_{i,t} < 0$ the exponential of multiplication factor is positive. And when $1/e \leq \frac{L}{e f_{i,t}} \leq 1$, then $0 < \prod_{i=1}^{n} \left(\frac{L}{e f_{i,t}}\right)^{\frac{n f_{i,t} - L}{L}} < 1$, so $\frac{dH_t}{dt} = ln \prod_{i=1}^{n} \left(\frac{L}{e f_{i,t}}\right)^{\frac{n f_{i,t} - L}{L}} < 0$, in this time H_t will has a progressive decrease with increasing time. When $\frac{L}{e f_{i,t}} > 1$ then $\prod_{i=1}^{n} \left(\frac{L}{e f_{i,t}}\right)^{\frac{n f_{i,t} - L}{L}} > 1$ so $\frac{dH_t}{dt} > 0$, in this time $H_t$increases progressively with increasing time. Similarlywhen $L - n f_{i,t} > 0 \frac{1}{e} \leq \frac{L}{e f_{i,t}} \leq 1 H_t$ is an increasing function of time; When $\frac{L}{e f_{i,t}} > 1$ H_t is a decreasing function with time. We can see from abovewhen $\begin{cases} L - n f_{i,t} < 0, \\ \frac{L}{e f_{i,t}} > 1, \end{cases}$ or $\begin{cases} L - n f_{i,t} > 0, \\ \frac{L}{e f_{i,t}} \leq 1, \end{cases}$ (n¿e) is that when $f_{i,t} < \frac{L}{n}$ or $\frac{L}{e} < f_{i,t} < L$the value of H_t will increase with increasing evolution. At the same timewhen $\frac{L}{e} < f_{i,t} < \frac{L}{e}$the value of$H_t$ will decrease with increasing evolution. So we can know that at the beginning of GEP, genes have influenced by random factors so that each symbol occurred mostly in the interval $\left[\frac{L}{n}, \frac{L}{e}\right]$and population genetic uniformity is higher than other time. In

Table 2: Parameters Sets of GEP & JM-GEP

Parameters	Span Solution	Parameters	Span Solution
Population Size	60	Function Set	$\{+-*/\wedge e^x\}$
Gene Num	5	Terminal Set	$\{x,0,1,\ldots 9\}$
Head Length	6	Select Operator	Roulette Wheel
Maximum of Generations	1000	Transposition Operator	0.1
Mutation Operator	0.044	Recombination Operator	0.3
Jumping Operator	JM-GEP with Equation(5)		
Fitness Function	GEP with Equation(2)(R=100), JM-GEP with(4)		
Terminal Set	Maximum of Generation		

this time, gene sequence entropy H_t is decreasing function with time. And because the population diversity is inversely correlated with H_t, the population diversity in this stage is increasing with t. But after jump operator joining in GEP, that is JM-GEP, JM-GEP can achieve its maximum diversity, and it cannot be attained at biological evolution. However, in order to get the optimal solution at later stage, population genetic is developing with same direction, the symbols have been distributed on the interval $[0,\frac{L}{n}]$and $[\frac{L}{e},1]$, and genetic uniformity is reduced, H_t is increasing function with t, at the same time population diversity is decreasing with$t$. Through the analysis of GEP and JM-GEP, when jump operator joins GEP, each generation of population diversity is superior than GEP, thus in the process of the whole evolution, JM - GEP is more quickly reaching a maximum values than GEP. In other word JM-GEP has accelerated the speed at the end of generation, the algorithm has higher and faster search ability.

3.2 JM-GEP Complexity Analysis

Theorem 1: the complexity of the algorithm is $O(P\times G\times n)$, where P is the population size, G is the maximum of generations, n is the sample size.

Proof: In the algorithm, the calculative complexity of population initialization from n samples is $O(n)$; each generation population needs to calculate, so complexity is $O(P\times n)$; as the maximum generation is G , so the complexity of the algorithm is $O(P\times G\times n)$.

3.3 JM-GEP Convergence Analysis

Theorem 2: JM-GEP convergence for optimal solution is as follows.

$$P_{Jm}=\frac{2}{k\cdot(kh-1)\cdot(h+3)\cdot h}\cdot P_{jm}$$

Proof: Individual C will be selected to jump when its probability is reach to P_{jm}. Start point of jumping is set as $f_{i,j}\quad i=1,2,\ldots k; j=1,2,\ldots h$, so that there exists kh jump factors. While, each jump factor location can be inserted has a total of $kh-1$ except itself. So each factor may produce the number of chromosome is $(kh-1)\cdot k(h-i-1)$ and use f_{ij} as a current position of jump factor. We can know from above, individual C may generate the number of the chromosome is $\sum_{i=1}^{n}(kh-1)\cdot k(h-i-1)=\frac{k\cdot(kh-1)\cdot(h+3)\cdot h}{2}$. Then jumping genes possible probability is $\frac{2}{k\cdot(kh-1)\cdot(h+3)\cdot h}\cdot P_{jm}$.

4 Heavy Metals prediction Modelling Based on GEP and JM-GEP

To illustrate the actual application process about heavy metals prediction modelling based on GEP and JM-GEP, we made model aim at Cr content and PH changed over time to prove its feasibility and effectiveness. The data series are the former 16 data of the experimental in Hebei University of Engineering. Finally, the simulation experimental results demonstrate the effectiveness and practicability of the proposed GEP and JM-GEP algorithm.

4.1 A model of Cr content changed over time based on GEP

The data series of Cr content changed over time are shown in table 1. Where t_i is changed time measured every 4 week. T_i is the content of Cr, is namely the next content. In this paper we have formed GEP model and JM-GEP model just on T_i. Parameters of the algorithms in the test are set as shown in Table 2. In the mixed programming environment of VC++6.0 and matlab7.2, run the evolutionary computation programming for the time sequence of in table 1. After 1000 generations of evolutionary computation, we get better adaptability models and structures expression as shown in following type (15) (16):

$$Y_{GEP}(x) = -6.780092 + \frac{x}{1.313761} \cdot 10^{0.402417} + 2x^2 \cdot 10^{0.126865} \tag{15}$$

$$Y_{JM-GEP}(x) = 1.1425 + 2x + 3x^2 + \frac{0.962584}{x} - \frac{x}{0.160436} - 10^{0.646596} \tag{16}$$

4.1.1 Model Emulation

Figure (10) and figure (11) shows the model (15) (16) of simulation picture of the heavy metals changed over time. Where, x-axis is the value of x(serial number of Cr changed over time) in table 1. Y-axis is T_i (the accumulated data of HM changed over time).

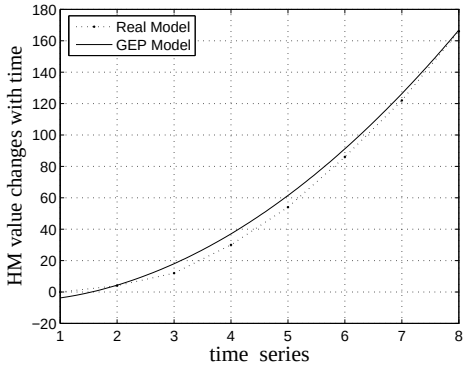

Fig. 10: Simulation Result of Model(15)

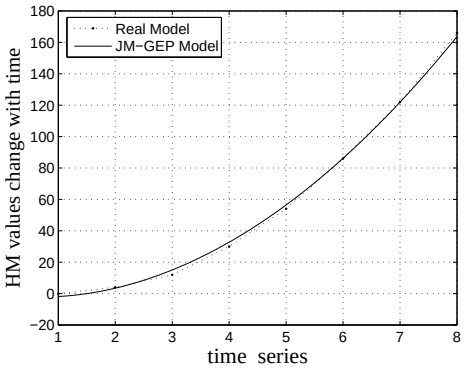

Fig. 11: Simulation Result of Model(16)

From figure10,11 we can see, GEP and JM-GEP are better matching the form of HM changed over time. But compared with these, in 10 GEP run the 350 generations when program has found optimal solution, and the best fitness is 1549.791070 and the time consume is 10 seconds. While in Figs.11 JM-GEP run the 100 generations when program has found the optimal solution, and the best fitness is 4747.393139 and time is only taken 4 seconds. This fully shows that JM - GEP model has higher prediction efficiency, and better fitness(Fitness value represents the error of the predicted values and the actual value, the error is smaller, the greater the adaptive value is). So improved HM prediction model based on JM-GEP is more ideal than GEP. We have formed GEP model and JM-GEP model on t_i. Structures of expression as shown in the following type (17) (18).

$$Y_{GEP}(x) = x + log|log|x - 0.941907|| + log|-0.656240 - 2x| + lnlogx^{1/2}sinx + 3.507761x \tag{17}$$

$$Y_{JM-GEP}(x) = \frac{x}{0.253578} - 1.31911 + x - logtan(10^x + x) + log\left|sintan(log\frac{e^x}{0.905942})\right| - tane^x + log|1.429801 - x| \tag{18}$$

Figure (12)(13) shown the model(17)(18) of the week number on HM changed. Simulation result(17)(18) indicates GEP and JM-GEP are better matching the week number of HM changed over time whether for Cr content changed over time or the week number sequence of HM . We use GEP model(15) and JM-GEP model(16) on Cr content changed over time to evaluate HMFT and JM-GEP.

4.1.2 Compared with other models

Table 3 is the evaluation results on GEP and JM - GEP compared with other models. You can see from table 3

GEP and JM - GEP can better simulate the data of heavy metals form changed over time, but JM - GEP have better simulation than GEP.

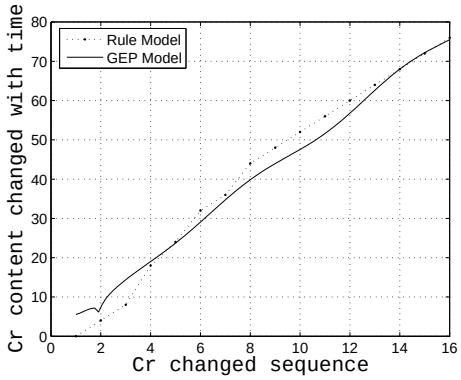

Fig. 12: Simulation Result of Model(17)

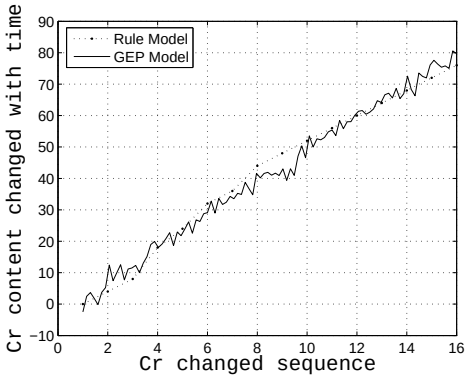

Fig. 13: Simulation Result of Model(18)

4.1.3 HMFT Parameters

By the model type (15), (16) calculated GEP cumulative time or next changed time(17th) of heavy metals is 227.4839, JM-GEP is 211.720357, the real result is 218. While, GEP in 17th interval changed time of HM is 73.4839, JM-GEP is 80.720357, and the real result is 80; New model JM-GEP measured values are smaller than GEP, so an proved HM prediction modelling based on JM-GEP is applicable. Table 4 is evaluation results in 17th point average interval time and the next changed time using several classic parametric model, non parameter model and GEP model.

4.1.4 Model Fitting Degree

Model fitting goodness can be used to predict evaluation models performance. This method is calculation model of the deviation between the predicted values and the real values, it can use X square test or Kolmogorov - Smimov test method to assess. This paper selects error square method to calculate. Supposed f is new model predicted value, y is real value of Cr content changed over time. So the definition of error is (19)

$$error = \sum_{i=1}^{m}(f_i - y_i)^2 \tag{19}$$

Where y is the number of data, m is total number of data. Using model fitting goodness formula (19) calculated the sum of error square between predicted values and real values with GEP model (15) and JM - GEP model (16), respectively. Then compared with GP and BP model, the detailed results as shown in table (6).

Table 6: Model Fitting Degree Evaluation

Evaluation Index	JM-GEP Model	GEP Model	GP Model	BP Model
Fitting Degree	3.469e +003	4.008e +003	2.687e +004	2.4397e +004

So, the approximation fitting degree of GEP model is good to Cr content changed over time. But JM - GEP model has a smaller error and higher precision. The fit of the two models are better than GP model, and the training of BP model.

4.2 Models of PH changed over time

4.2.1 Modelling and Simulation

Table 6 is data series of PH changed over time(only 16 data).T_i is the number of weeks of HM changed(here average every 4 weeks measured), t_i refers to the PH value of heavy metals. We have formed GEP and JM-GEP model on t_i and T_i .Parameters as shown in table 2. In the mixed programming environment of VC++6.0 and matlab7.2, run the evolutionary computation programming for the time sequence of in table 6. After 1000 generations of evolutionary computation, we get better adaptability models and structures expression as shown in the following type (20) (21):

$$Y_{GEP}(x) = 3.766598 - x_2^{\frac{1}{3}} + \sqrt{x_2} + (\frac{\sqrt{x_1}}{0.504074})^{\frac{1}{8}} + 10 \cdot (\frac{0.747917}{\sqrt{x_1}})^{\frac{1}{4}} \tag{20}$$

Table 3: Evaluation Results of data in table 1

Evaluation Model	t_i	Next Changed Value	Evaluation Model	t_i	Next Changed Value
GEP Model	63.4839	227.4839	G-O Model	70.2176	234.684690
JM-GEP Model	45.720357	211.720357	Moranda Model	68.3674	242.400123
Exponential Model	64.5823	251.879013	S-W Model	71.547330	249.569820
J-M Model	66.9940	267.505001			

Table 4: The results of HMFT

Evaluation Models	HMFT	Next changed values	Evaluation Models	HMFT	Next changed values
JM-GEP Model	45.720357	211.720357	U-M Model	26.75	287.75
GEP Model	63.4839	227.4839	G-O Model	83.4742	424.4742
GP Model	35.4307	229.7812	S-W Model	126.7990	442.7990
BP Model	41.0000	300.0000	J-M Model	108.5019	469.5019

Table 5: Data series of PH changed over time

x	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
t_i	0	4	8	18	24	32	36	44	48	52	56	60	64	68	72	76
T_i	12.2	11.7	11.6	11.3	10.5	10.4	9.6	8.7	8.5	8.8	9.6	9.11	9.5	10.05	10.03	11.01

$$Y_{JM-GEP}(x) = 9.816278 + x_2^{\frac{1}{4}} - x_1 + \sqrt{x_1} + 10^{10^{-0.063631-x_1}} \quad (21)$$

Model simulation structure as shown in figure (14),(15).

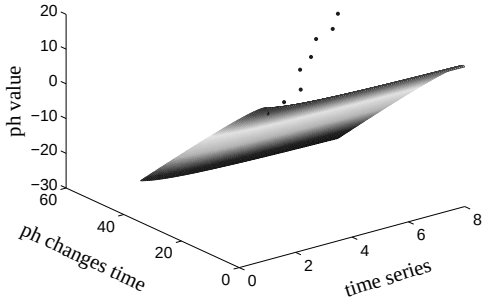

Fig. 14: Simulation Result of Model(20)

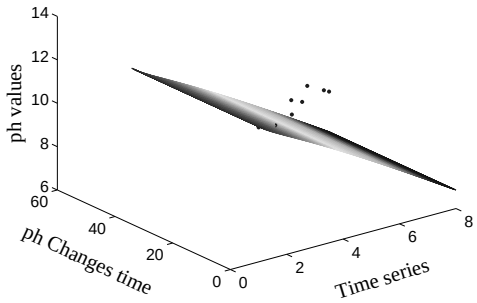

Fig. 15: Simulation Result of Model(21)

4.2.2 HMFT Parameters

We have also calculated the PH value of the next time ,in other words, PH value at 17th for GEP is 9.20157 and JM-GEP is 6.9015 in the model (20) (21) , while, the real value is 7.1 . Accordingly, the average weekly number

t_{GEP} is 83.074, t_{JM-GEP} is 80.50, and the real measured result is 80.Clearly,thenew model simulation value and design value less than the difference, so it is applicable to this software system. Table 7 is evaluation values using classic parametric model, several typical non parameter model and GEP model on PH changed over time of heavy metals by 17 points.

Table 7: Evaluation Results of 17th PH changed over time

Evaluation Models	Next Changed Values	Evaluation Models	Next Changed Values
GEP Model	9.20157	G-O Model	10.5427
JM-GEP Model	6.9015	Moranda Model	11.6504
Exponential Model	11.6023	S-W Model	13.4530
J-M Model	12.4301		

Calculated by the table (7), the data can be seen that several kinds of parametric model simulation values and design values more than the difference under next PH changed over time, and G - O model, Moranda model shows a better value, but still not as good as the GEP model. While new model JM-GEP calculation results are ideal, predicted the next changed time is closer than the actual value. This also verified the new model is better than other several prediction ability.

4.2.3 Model Fitting Degree

Fitting degree formula (19) calculated the error sum of squares between predicted values and the real value with GEP model (20) and JM - GEP model (21), GP model, BP model respectively, the detailed results as shown in table(8). Model Fitting Degree Evaluation

Table 8: Model Fitting Degree Evaluation

Evaluation Index	JM-GEP Model	GEP Model	GP Model	BP Model
Fitting Degree	3.5826e +003	2.0940e +003	4.5690e +003	4.6743e +002

So, the approximation fitting degree of GEP model is good to HM PH changed over time. But JM - GEP model has a smaller error and higher precision than GEP. The fit of the two models are better than GP model and the training BP model.

5 Conclusions

With GEP forecast, we not only do not need to know the causal relationship among the various factors, but also not to know the objective function, only need to provide enough experiments or the experimental data, can achieve the purpose of accurate prediction. In this paper, the main idea is to get the mathematical model of a better prediction precision and fitting degree for the time series about morphological of HM changed over time based on JM-GEP. This method is more super than the traditional method of GEP and also faster than GEP on solving problems.

Acknowledgement

The authors are grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Quan G, Yan J. Binding constants of lead by humic and fulvic acids studied by anodic stripping square wave voltammetry [J]. Russian Journal of Electrochemistry, 2010, 46(1): 90-94.
[2] Christl I, Metzger A, Heidmann I, et al. Effect of humic and fulvic acid concentrations and ionic strength on copper and lead binding [J]. Environmental Science and Technology, 2005, 39 (14):5319-5326.
[3] Hua Zhang, Pinjing He, Fan Lv, et al. The Research Progress of Chemical Speciation Analysis of Heavy Metals In The Environment [J]. Environmental chemistry, 2011, 30(1): 130-137.
[4] Weng L, Temminghoff E J M, Van Riemsdijk W H. Contribution of individual sorbents to the control of heavy metal activity in sandy soil [J]. Environmental Science and Technology, 2001, 35(22): 4436-4443.
[5] Hiemstra T, van Riemsdijk W H. A surface structural approach to ion adsorption: the charge distribution (CD) model [J]. Journal of Colloid and in Terface Science, 1996, 179(2): 488-508.
[6] Dijkstra J J, Meeussen J C L, Comans R N J. Leaching of heavy metals from contaminated soils: an experimental and modelling study [J].Environmental Science and Technology, 2004, 38(16): 4390-4395.
[7] Anirhomayoun Saffarzadeh, Takayuki Shimaoka. Impacts of natural weathering on the reansformation/neiformation processes in landfilled MSW bottom ash: A geoenvironmental perspective[J]. Waste Management, 2011, 31(12):2440-2454.

[8] Yi Wai Chiang, karel Ghyselbrecht, Rafael M. Santons. Sythesis of zeolitic-type adsorbent material from municipal solid waste incinerator bottom ash and its application in heavy metal adsorption[J].Catalysis Today, 2012, 190(1): 23-30.
[9] M.Gori, B.Bergfeldt, G. Pfrang-Stotz. Effect of short-term natural weathering on MSWI and wood waste ash leaching behaviour[J]. Waste Management, 2011, 189(1-2):435-443.
[10] Ferreira C. Gene expression programming: A new adaptive algorithm for solving problems[J].Complex Systems,2002,12(2):87-129.
[11] Angeline P J. Adaptive and Self-Adaptive Evolutionary Computation, Piscataway: IEEE Press, 1995,152-163.
[12] Hinterding R, Michalewicz Z, Eiben A E. Adaptation in Evolutionary Computation: A Survey[A]. Proceedings of the 4th IEEE Conference Evolutionary Computation[C]. Indianapolis, USA: IEEE Press, 1997, 65 69.
[13] Eiben A E, Hinterding R, Michalewicz Z. Parameter Control in Evolutionary Algorithms[J]. IEEE Transactions on Evolutionary Computation, 1999, 3(2): 124 141.
[14] Ishibuchi H, Shibata Y. Mating Scheme for Controlling the Diversity-Convergence Balance for Multiobjective
[15] Shimodaira H. A Diversity Control-oriented-genetic Algorithm (DCGA): Performance in Function Optimization[A]. IEEE Congress on Evolutionary Computation (CEC2001)[C]. Seoul, Korea: IEEE Press, 44 51.
[16] Musa J D. Software Reliablity Engineering. New York: Mc Graw Hill. 1999.
[17] Ansheng Chen, Zhihua Cai, Qiong Gu, Liechao Zhang. A New GEP Algorithm and Applications[J]. Computer Research of Computers, 2007,24(6):98-103.
[18] Jie Zuo. The Core Technology Research of Gene Expression Programming [D]. SichuanSichuan university, Ph.D. Thesis, 2004.
[19] Jing Xiao. Research and Application of Gene Expression Programming in Software Reliability Modelling[D]. Hebei: Hebei University of Engineering, M.D. Thesis, 2012.
[20] Shannon C E. A Mathematical Theory of Communication[J]. Bell System Technical Journal,1948,27:379-423,623-656.
[21] Jaynes E T. Information Theory and Statistical Mechanics[J].Physical Review,1957,106(4):620-430.
[22] Xiaolong Wang, Zhifa Yuan, Mancai Guo. Maximum information entropy principle and genetic balance[J]. Journal of Genetics and Genomics,2002,29(6):562-564.
[23] Changan Yuan, Yuzhong Peng, Xiao Tan. Gene expression programming algorithm principle and application[M]. Beijing: Science press,2010.

Yongqiang ZHANG Professor of Hebei University of Engineering. His research interests are Gene Expression Programming Algorithm and software reliability engineering.

Junxia LI candidate for master degree who is studying on GEP Algorithm and the heavy metals prediction modeling.