ISSN 1330-3651 (Print), ISSN 1848-6339 (Online)
DOI: 10.17559/TV-20160525104127

REGRESSION MODELLING BASED ON IMPROVED GENETIC ALGORITHM

Shi Minghua, Xiao Qingxian, Zhou Benda, Yang Feng

Original scientific paper
Regression model is a well-established method in data analysis with applications in various fields. The selection of independent variables and mathematically transformed in a regression model is often a challenging problem. Recently, some scholars have used evolutionary computation to solve this problem, but the result is not effective as we desired. The crossover operation in GA is redesigned by using Latin hypercube sampling, then combining two commonly used statistical criteria (AIC, BIC) we are presenting an improved genetic algorithm based for solving statistical model selection problem. The proposed algorithm can overcome strong path-dependence and rely on experience of classical approaches. Comparison of simulation results in solving statistical model selection problem with this improved GA, traditional genetic algorithm and classical algorithm for model selection show that the new GA has superiority in solution of quality, convergence rate and other various indices.

Keywords: *genetic algorithm; Latin hypercube sampling; regression analysis; regression model selection*

Regresijsko modeliranje zasnovano na poboljšanom genetičkom algoritmu

Izvorni znanstveni članak
Regresijski model je dobro uhodana metoda u analizi podataka s primjenom u raznim područjima. Izbor nezavisnih varijabli i matematički transformiranih u regresijski model, često predstavlja izazovan problem. Nedavno je nekoliko znanstvenika primijenilo evolucijski proračun za rješenje tog problema, ali rezultat nije učinkovit onoliko koliko smo željeli. Ukrižena (crossover) operacija u GA redizajnirana je primjenom Latin hypercube uzorkovanja, a zatim, kombinacijom dvaju uobičajeno korištenih statističkih kriterija (AIC, BIC), dajemo poboljšani genetički algoritam za rješavanje problema izbora statističkog modela. Predloženim se algoritmom može prevladati jaka ovisnost o putanji i osloniti na iskustvo stečeno primjenom klasičnih pristupa. Usporedba rezultata simulacije u rješavanju problema odabira statističkog modela s ovim poboljšanim GA, tradicionalnog genetičkog algoritma i klasičnog algoritma za odabir modela pokazuje da je novi GA superiorniji u rješavanju kvalitete, brzine konvergencije i drugih različitih pokazatelja.

Ključne riječi: *genetički algoritam; Latin hypercube uzorkovanje; odabir regresijskog modela; regresijska analiza*

1 Introduction

Regression analysis is an important statistical method describing the mutual dependence of several variables. Due to its higher prediction and control capacity, this process is widely applied in natural science, social science and engineering. Regression analysis involves 3 major aspects: selection of major variables from a myriad of explanatory variables (variable selection); selection of an appropriate function to describe the model (transformation); and parameter estimation of the regression model. The commonly used approaches include stepwise regression, expert's selection and complete model method [1, 2]. All these methods share two common weak points: (1) Heavy dependence on the subjective experience of the researchers [1]; (2) Complex and time-consuming process dealing with high-dimensional data. The reason is that every possible subset of variables has to be tested when selecting the variables. For instance, there will be $2m$ possible subsets for m variables, and the calculation load is huge when m is large.

As a representative of the intelligent optimization algorithms, genetic algorithm is highly efficient and bionics-based and commonly applied to optimization problems in various fields. Possessing the features of self-learning, self-organization and self-adaptation as well as simplicity, universality and robustness, genetic algorithm is suitable for parallel processing and optimization in large space. Genetic algorithm is also a typical method for solving the regression model. A number of methods for regression modeling based on intelligent optimization algorithms have emerged recently, including those based on genetic algorithm [3], particle swarm algorithm [4], ant colony algorithm [5], neural network [6], and simulated annealing algorithm [7].

The crossover operator used for genetic algorithm is redesigned based on Latin hypercube sampling (LHS), and a method for regression modeling based on improved genetic algorithm is proposed. Compared with the method in literature [3 ÷ 7], our algorithm has the following features: (1) the model is driven by data, with parallel implementation of variable selection, transformation and parameter estimation. Most of the existing algorithms can either deal with either variable selection or parameter estimation or both, but very few of them can achieve transformation. Among the algorithms that consider transformation, the focus is placed on which transformation variable selection and parameter estimation can be implemented. This result is in strong dependence on the sequence of processing [8]; (2) We adopt the generalized linear regression model which includes transformation itself and a wider application scope than the simple linear model in literature [3 ÷ 7]; (3) Using the crossover operator redesigned by LHS, the proposed algorithm is capable of building regression model on high-dimensional data with a large sample size. In contrast, the regression model is built in literature [3 ÷ 7] by using simple genetic algorithm directly without any necessary improvement; (4) It has been proven theoretically that LHS is superior to uniform design sampling and good-point set sampling

2 LHS and its properties

2.1 LHS

For a deterministic function $Y = f(X)$, where $Y \in R$

$X \in R^p$, p statistically independent elements $X_1, X_2,\ldots, X_p$ of random vector $\boldsymbol{X}$ are taken as the input variables. In order to estimate the mean of response variable Y using fewer samples, Mckay et al. put forward Latin hypercube sampling (LHS) in 1979 [10]. Owen pointed out that LHS was a technique comparable to jittered sampling. He demonstrated that LHS had a faster convergence than Monte Carlo sampling and the sampling points generated by LHS represented all parts of the design space. Moreover, there was no need to bother with dimensionality, and the number of samples can be any integer [13]. LSH was widely used to improve the algorithm efficiency, such as Cholesky decomposition [12], single switch optimization algorithm [13], Columnwise-pairwise algorithm [14], Rank Gram-Schmidt algorithm [15] and simulated annealing algorithm [16].

Literature [17] believed that searching for better samples in the ancestors with high fitness was crucial for the efficiency improvement of genetic algorithm. On this basis, the good point genetic algorithm (GGA) was proposed. The good point set of n points over the t-dimensional cube is $[0,1]^t$ $p_n(i) = \{(\{r_1\times i\}, \{r_2\times i\}, \{r_3\times i\},..., \{r_t\times i\}), i = 1, 2,..., n\}$ where $r_k = 2\cos(2\pi k/p)$, $1 \le k \le t$, and p is the minimal prime larger than $2t+3$; or $r_k = e^k$, $1 \le k \le t$, where $\{a\}$ denotes the decimal part of a. It is obvious that once n is known, the good point set is also determined.

The point distribution of the good point set has directionality but no randomness. Therefore many lattices are skipped, and the corresponding points will not be selected as the descendents of crossover, which affects the overall searching effect. Based on literature [17], we introduce LHS into the genetic algorithm.

For the t-dimensional cube $[0,1]^t$, LHS with n points is implemented in the following steps:

1. The interval of coordinates [0,1] in each dimension is equally divided into n parts. Label i denotes the smaller interval $\left(\frac{i-1}{n}, \frac{i}{n}\right]$, and $(\pi_{1j}, \pi_{2j},..., \pi_{nj})'$ one random arrangement of n labels (1, 2,..., n) of coordinates in the j^{th} dimension.
2. Suppose that the t random arrangements are mutually independent, and the random matrix $\pi = (\pi_{ij})_{n\times t}$ of order $n\times t$ is obtained;
3. Let

$$c_{ij} = \frac{\pi_{ij} - \frac{1}{2} + u_{ij}}{n}, i = 1,..., n, j = 1,..., t \quad (1)$$

where u_{ij} is the sample obeying independent indentity distribution on $[-\frac{1}{2}, \frac{1}{2}]$ within dependence from π.

2.2 Properties of LHS

Sampling quality can be measured by using standard deviation, histograms and means, but these indicators are subjective, which cannot fully depict overall distribution. Hence some literature [18, 19] used bias and L_2 bias as more accurate measure. However, Hickernell [20] pointed out that bias and L_2 bias were not suitable as general indicators and demonstrated the reasonability of using WD bias to characterize LHS and random uniform design. However, no WD bias calculation was provided for a specific case of sampling. Here we calculate WD bias for LHS and random uniform design and use it as an indicator for comparing the two sampling methods:

$$WD(P) = \left\{-\left(\frac{4}{3}\right)^s + \frac{1}{n^2}\sum_{k=1}^{n}\sum_{j=1}^{t}\prod_{i=1}^{s}\left[\frac{3}{2} - \left|x_{ki} - x_{ji}\right|\left(1 - \left|x_{ki} - x_{ji}\right|\right)\right]\right\}^{\frac{1}{2}}$$

where $x_k = (x_{k1}, x_{k2}, \cdots, x_{ks})$; n is the number of samples, s is the dimensionality of the samples; $P = \{x_1, x_2, \cdots, x_n\}$ is the point set sampled over $C^s = [0,1]^s$ cube.

Theorem 1 Let the elements of $L_{n,s}$ be latin hypercube samples. Then we have the following:

$$E(WD(L_{n,s})^2) = -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\left(\frac{4}{3} + \frac{1}{6n^2} - \frac{1}{6n}\right)^s$$

Prove:

$$E(WD(L_{n,s})^2) = -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\left(\frac{3}{2} - \sum_{i=1}^{n}\sum_{j=1}^{n}\frac{|i-j|}{n}\left(1-\frac{|i-j|}{n}\right)P(i,j)\right)^s =$$

$$= -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\left\{\frac{3}{2} + \frac{2}{n(n-1)}\sum_{i>j}\left[\frac{(i-j)^2}{n^2} - \frac{i-j}{n}\right]\right\}^s =$$

$$= -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\cdot\left[\frac{3}{2} + \frac{2}{n(n-1)}\cdot\frac{(n-1)(1-n-n^2)}{12n}\right]^s =$$

$$= -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\cdot\left(\frac{3}{2} + \frac{1-n-n^2}{6n^2}\right)^s =$$

$$= -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\cdot\left(\frac{3}{2} + \frac{1}{6n^2} - \frac{1}{6n} - \frac{1}{6}\right)^s =$$

$$= -\left(\frac{4}{3}\right)^s + \frac{1}{n}\left(\frac{3}{2}\right)^s + \left(1-\frac{1}{n}\right)\cdot\left(\frac{4}{3} + \frac{1}{6n^2} - \frac{1}{6n}\right)^s$$

Theorem 2 Let the elements of $R_{n,s}$ be random uniform design samples. Then we have the following:

$$E(WD(R_{n,s})^2) = \frac{1}{n}\left[\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s\right]$$

Prove:

$$E(WD(R_{n,s})^2) = \frac{1}{n}\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s + \left(1-\frac{1}{n}\right)\left(\int_0^1\int_0^1\left(\frac{3}{2} - |x-y|\left(1-|x-y|\right)\right)\mathrm{d}x\mathrm{d}y\right)^s =$$

$$= \frac{1}{n}\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s + \left(1-\frac{1}{n}\right)\left(\int_0^1 \mathrm{d}y\int_0^y\left(\frac{3}{2} - (y-x)\left[1-(y-x)\right]\right)\mathrm{d}x + \right.$$

$$\left. + \int_0^1 \mathrm{d}y\int_y^1\left(\frac{3}{2} - (x-y)\left[1-(x-y)\right]\right)\mathrm{d}x\right)^s =$$

$$= \frac{1}{n}\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s + \left(1-\frac{1}{n}\right)\left\{\int_0^1\left(\frac{3}{2}y + \frac{1}{2}(y-x)\Big|_0^y - \frac{1}{3}(y-x)^3\Big|_0^y\right)\mathrm{d}y + \right.$$

$$\left. + \int_0^1\left(\frac{3}{2}(1-y) + \frac{1}{2}(x-y)^2\Big|_y^1 + \frac{1}{3}(x-y)^3\Big|_y^1\right)\mathrm{d}y\right\}^s =$$

$$= \frac{1}{n}\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s + \left(1-\frac{1}{n}\right)\left\{\int_0^1 \left(\frac{3}{2}y - \frac{1}{2}y^2 + \frac{1}{3}y^3\right)\mathrm{d}y + \right.$$

$$\left. + \int_0^1 \left[\frac{3}{2}(1-y) + \frac{1}{2}(x-y)^2 + \frac{1}{3}(x-y)^3\right]\mathrm{d}y\right\}^s =$$

$$= \frac{1}{n}\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s + \left(1-\frac{1}{n}\right)\left(\frac{3}{4} - \frac{1}{6} + \frac{1}{12} + \frac{3}{4} - \frac{1}{6} + \frac{1}{12}\right)^s =$$

$$= \frac{1}{n}\left[\left(\frac{3}{2}\right)^s - \left(\frac{4}{3}\right)^s\right]$$

Property 1 the WD bias of sample point set in LHS is smaller than that in random uniform design (n>1).

Prove:

$$E(WD(R_{n,s})^2) - E(WD(L_{n,s})^2) = \left(1-\frac{1}{n}\right)\left[\left(\frac{4}{3}\right)^s - \left(\frac{4}{3} + \frac{1}{6n^2} - \frac{1}{6n}\right)^s\right] > 0$$

A small WD bias indicates that the sample points have a more uniform distribution in the sample space and contain the information of the entire space. Therefore, the model built on these sample points can more accurately capture the trend and variation of the function [20]. As known from property 1, the WD bias of sample point set in LHS is smaller than that in random uniform design. Therefore, LHS is superior to random uniform design. Moreover, literature [21] has proved both theoretical and experimentally that random design sampling is superior to good-point set sampling, so we infer that LHS is superior to good-point set sampling.

3 Solving the regression model

3.1 Problem description

A classical regression problem is described as follows: According to a certain judgment rule, $p(p \le m)$ independent variables are selected from m variable sets $X = \{X_1, X_2, \cdots, X_m\}$ so as to build the optimal regression model against the dependent variable Y :

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1' + \hat{\beta}_2 X_2' + \cdots + \hat{\beta}_p X_p'$$

where

$$X' = \left\{X_1', X_2', \cdots, X_p'\right\} \subseteq X, \ \left\{\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_p\right\}$$

is the least square estimate.

To further improve the precision of the regression model, Cook and Weisberg [20] pointed out the need for transformation using function with respect to X'. They described the regression model as follows: choose an appropriate subset of independent variable X' and function T that make

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 T(X_1') + \hat{\beta}_2 T(X_2') + \cdots + \hat{\beta}_p T(X_p')$$

optimal under certain model selection rule.

3.2 Information criteria for the statistical model

Model selection rule has long been a research focus in statistics. Among various, we choose two representatives. One is Akaike information criterion(AIC) based on the distance between the real model and the candidate model using Kllback-Leibler information. For the regression model, $AIC = \log(S_p^2) + 2p/n$, where n is the number of samples; p is the number of independent variables in the regression equation; S_p^2 is the residual variance. The other is the Bayesian information criterion(BIC) proposed by Schwarz assuming that the candidate model family has a uniform distribution. Hence the model with the maximum posterior probability is selected. For the regression model, $BIC = \log(S_p^2) + (p/n) \times \log n$.

3.3 Encoding and fitness function

The chromosomes are expressed by two-stage real-number encoding (T,E), where T denotes whether the variable is selected and the transformation adopted after selection. E is the exponent of the function. Three forms of functions are chosen for the transformation, namely, power function, logarithmic function and exponential function. In fact, other forms of functions can be used as well depending on the specific needs. If $T = \{0, 1, 2, 3\}$, $E = \{-6, -5, -4, -3, -2, -1, 1, 2, 3, 4, 5, 6\}$, $t_i \in T$, $e_i \in E$, Then

$t_i = 0$: $T(X_i) = 0$, indicating this variable is not included;

$t_i = 1$: $T(X_i) = (X_i)^{\frac{e_i}{2}}$;

$t_i = 2$: $T(X_i) = (\ln X_i)^{\frac{e_i}{2}}$;

$t_i = 3$: $T(X_i) = (e^{X_i})^{\frac{e_i}{2}}$.

The smaller the AIC and BIC, the higher the quality of the model will be. Thus the fitness function for model selection problem is assumed as the reciprocal of the rule function.

3.4 LHS crossover operator

Roulette-wheel selection is used for selecting 2 chromosomes, which are $A_1 = (a_1^1, a_2^1, \cdots, a_i^1, \cdots, a_l^1)$ and $A_2 = (a_1^2, a_2^2, \cdots, a_i^2, \cdots, a_l^2)$, respectively. If $i \le 4$, then $a_i^k \in T$; if $4 < i \le 16$, then $a_i^k \in E$, $k = 1, 2$. Let $\boldsymbol{H} = \left\{(x_1, x_2, \cdots, x_l) \,\middle|\, i \in J, x_i = *;\ i \notin J,\ x_i = a_i^1\right\}$, where $J = \{i \,|\, a_i^1 \ne a_i^2\}$ is the set composed of the loci of alleles A_1 and A_2, and let $n(J) = t$. Then using the t-dimensional curb composed of the loci of aleles, n points are selected for crossover by LSH.

Let the k-th chromosome in the n-th descendent be $b^{(k)} = \left(b_1^k, b_2^k, ..., b_l^k\right)$, where

$$b_i^k = \begin{cases} a_i^1 & i \notin J \\ \frac{1}{2}\langle c_{ij} \rangle & i = t_j \in J, 1 \le j \le t \end{cases}, \quad 1 \le k \le n, 1 \le i \le l$$

If $i \le 4$, then $b_i^k \in T$, and $\langle a \rangle$ indicates that a is mapped from interval [0,1] to the set {0,1,2,3} with squaring; if $4 < i \le 16$, then $b_i^k \in E$, and $\langle a \rangle$ indicates

that a is mapped from interval [0, 1] to set {−6, −5, −4, −3, −2, −1, 1, 2, 3, 4, 5, 6}. Specifically, let $x_{\max} = 6$, $x_{\min} = -6$ and y be the encoding after mapping. The mapping function is:

$$y_1 = x_{min} + (x_{\max} - x_{min}) \times x, \quad y = y_1 + \frac{\text{sgn}(y_1)}{2}$$

where $\text{sgn}(y_1)$ is the sign function of y_1. Then squaring is performed for y. Then the fitness values of n descendents are compared, and the chromosome (or chromosomes) with the largest value is chosen as the descendent of the crossover.

4 Performance testing of the improved genetic algorithm

The standard example provided by UCI was run in MATLAB 2010a, and a comparison was made with stepwise regression, complete model method and random genetic algorithm (RGA) [21]. The sample datasets were divided into two parts. One part was used for model construction (for large-scale sample size, the proportion of data in this part accounted for 80 %, or for 90 ÷ 95 % if the sample size was smaller). The other part was used for verification.

4.1 Validity analysis

The optimal criterion value for each run was recorded for each algorithm, and the optimal criterion values were compared after 150 runs. The minimal value was considered the best and denoted as Bestval, and the maximal value as the worst and denoted as Maxval. The time taken to find the optimal solution was also recorded for each algorithm, and the average of 150 runs was regarded the average time of optimum seeking and denoted as Besttime. Moreover, average criterion value $\overline{f}_{run}$ and optimal variance ρ_{run} in continuous runs were calculated.

Definition [25] Let

$$\overline{f}_{run} = \frac{1}{r}\sum_{j=1}^{r} f_j^*(T) \quad \text{and} \quad \rho_{run} = \frac{1}{r}\sqrt{\sum_{j=1}^{r}(f_j^*(T) - \overline{f}_{run})^2},$$

where $f_j^*(T)$ is the optimum obtained by genetic algorithm after T iterations for the j-th time; ρ_{run} is the standard deviation of the optimum recorded in r runs. However, the results are not reliable after a single run since the genetic algorithm contains many random operations, and the data can be easily contaminated. Therefore, the indicators of continuous runs are more important. For example, overall performance of the algorithm is measured by $\overline{f}_{run}$, and the algorithm stability by ρ_{run}.

As shown by the Tab. 1 and 2, the optimal criterion value of the intelligent optimization algorithm is superior to that of complete model method and stepwise regression. Moreover, the worst criterion value is also better. Since GGA cannot search the entire solution space, the criterion value of GGA is worse than that of RGA and LHGA. LHS, in contrast, has a higher sampling efficiency and ensures that the entire space is covered by the sample points. This is a major reason for the higher precision of LHGA compared with RGA. As shown in Table 1 and 2, the average criterion value and optimal variance in continuous runs of LHGA are smaller than those of RGA and GGA, indicating a better overall performance. The probability of approaching the optimum and the stability are both higher in each run with LHGA.

Table 1 Comparisons between LHGA and the other 2 algorithms, AIC are considered

Problem	Alg	Maxval	Bestval	$\overline{f}_{run}$	ρ_{run}	Besttime	Complete Model	StepwiseRegression
Heart Disease Index	*GGA*	17,0532	16,9947	17,0155	0,0108	58,3498	17,3402	17,2479
	RGA	17,0452	16,9745	17,0057	0,0157	49,8660		
	LHGA	17,0123	**16,9711**	**17,0020**	**0,0082**	**43,6295**		
Weather Ankara	*GGA*	2,2932	2,0909	2,2126	0,2294	304,5379	2,3449	2,345
	RGA	2,2105	2,0400	2,0809	0,1394	273,1118		
	LHGA	**2,1817**	**1,9050**	**1,9656**	**0,1343**	**249,1413**		
Housing	*GGA*	6,7865	6,4994	6,5909	0,0731	90,1297	8,2210	8,2329
	RGA	6,5570	6,4658	6,5139	0,0311	75,8487		
	LHGA	**6,5573**	**6,4445**	**6,4904**	**0,0154**	**68,7977**		
Concrete Compressive Strength	*GGA*	8,7678	8,5596	8,6069	0,0302	43,7629	10,080	10,0715
	RGA	8,6159	8,5553	8,5667	0,0137	40,7920		
	LHGA	**8,5967**	**8,3231**	**8,3334**	**0,0133**	**40,0896**		
Concrete Slump Test	*GGA*	3,6208	2,8244	3,1901	0,1713	17,7092	4,5396	4,4868
	RGA	3,0040	2,7887	2,9029	0,0361	18,4451		
	LHGA	**3,0045**	**2,6841**	**2,8068**	**0,0342**	**14,5150**		
Abalone (male)	*GGA*	4,7387	4,6401	4,6808	0,0223	353,1539	4,7640	4,7623
	RGA	4,6593	4,6288	4,6651	0,0068	331,0419		
	LHGA	**4,6527**	**4,3632**	**4,4689**	**0,0047**	**120,8126**		
Parkinsons Telemonitoring	*GGA*	8,1447	7,9867	8,1849	0,0216	5820,3	8,3213	8,2943
	RGA	8,0857	7,8491	8,0081	0,0077	6840,9		
	LHGA	**8,0660**	**7,7498**	**7,9542**	**0,0034**	**4980,7**		

Table 2 Comparisons between LHGA and the other 2 algorithms, BIC are considered

Problem	Alg	Maxval	Bestval	$\overline{f}_{run}$	ρ_{run}	Besttime	Complete Model	StepwiseRegression
Heart Disease Index	*GGA*	17,2041	17,1126	17,1629	0,0248	54,0396	17,5341	17,2959
	RGA	17,2217	17,0917	17,1453	0,0303	51,8489		
	LHGA	**17,1896**	**16,9912**	**17,0357**	**0,0209**	**40,8431**		
Weather Ankara	*GGA*	2,2759	2,1175	2,2185	0,1445	289,6587	2,3817	2,3778
	RGA	2,2436	2,0769	2,1074	0,1001	278,2713		
	LHGA	**2,1382**	**2,0694**	**2,0809**	**0,0090**	**239,4821**		
Housing	*GGA*	6,8778	6,6027	6,6955	0,0586	86,4383	8,3359	8,3305
	RGA	6,6827	6,5638	6,6312	0,0349	81,0427		
	LHGA	**6,6819**	**6,4687**	**6,5709**	**0,0178**	**70,2976**		
Concrete Compressive Strength	*GGA*	8,8352	8,6108	8,6580	0,0324	42,9781	10,1195	10,1013
	RGA	8,6555	8,5965	8,6137	0,0137	40,0047		
	LHGA	**8,6437**	**8,4031**	**8,4134**	**0,0124**	**38,4765**		
Concrete Slump Test	*GGA*	3,7966	3,1101	3,4361	0,1584	18,6621	4,5466	4,5016
	RGA	3,3099	3,0744	3,1823	0,0451	17,4561		
	LHGA	**3,3109**	**2,8585**	**2,9653**	**0,0410**	**14,0815**		
Abalone (male)	*GGA*	4,7592	4,6702	4,7095	0,0176	322,1194	4,7949	4,7870
	RGA	4,7012	4,6640	4,6764	0,0076	303,5799		
	LHGA	**4,6873**	**4,4632**	**4,4689**	**0,0055**	**112,9978**		
Parkinsons Telemonitoring	*GGA*	8,1659	8,0777	8,1034	0,0224	6649,4	8,3435	8,3095
	RGA	8,1037	8,0006	8,0037	0,0091	6531,1		
	LHGA	**8,1034**	**7,8664**	**7,9776**	**0,0013**	**5240,3**		

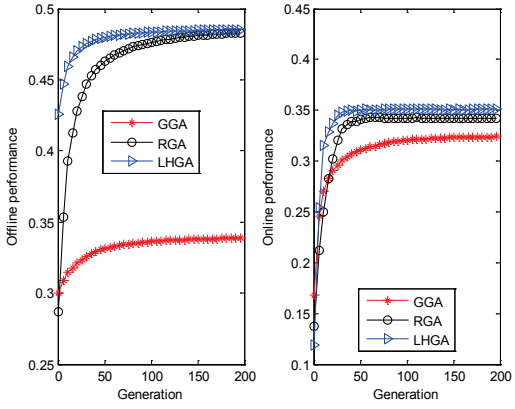

Figure 1(a) Weather Ankara performance result

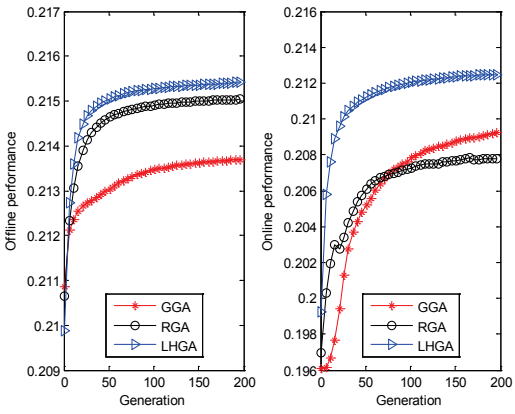

Figure 2(a) Abalone (male) performance result

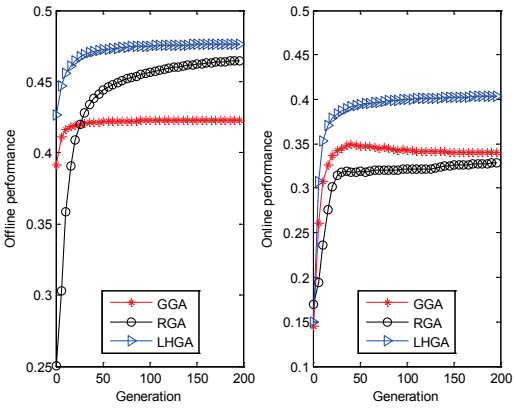

Figure 1(b) Weather Ankara performance result

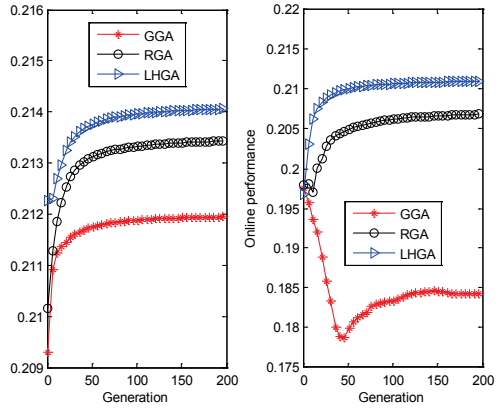

Figure 2(b) Abalone (male) performance result

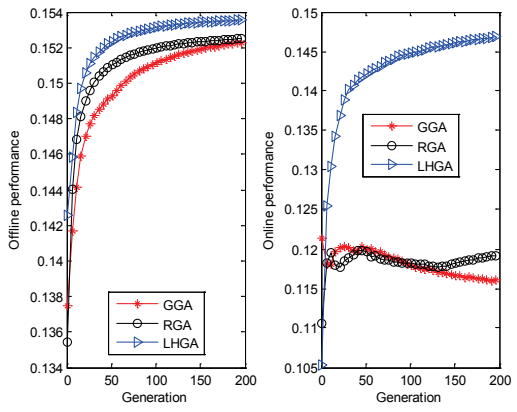

Figure 3(a) Housing performance result

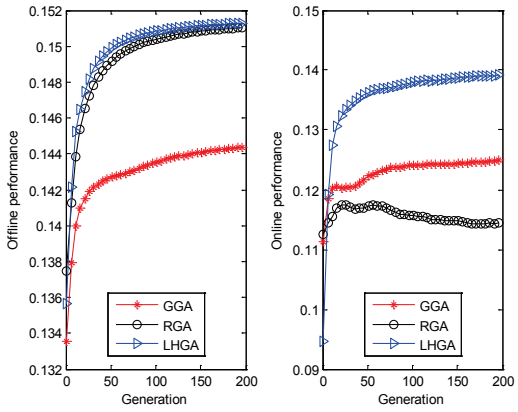

Figure 3(b) Housing performance result

Table3 The MSE of the forecasting data

Problem	Alg	AIC		BIC		Complete Model	StepwiseRegression
		AVG	VAR	AVG	VAR		
Heart DiseaseIndex	*GGA*	46,0316	0,1234	46,2375	0,1822	50,1256	49,0627
	RGA	45,9558	4,5706	46,1268	3,9660		
	LHGA	**43,8311**	**0,0631**	**44,0720**	**0,0767**		
Weather Ankara	*GGA*	1,5581	0,0333	1,6361	0,0522	1,8754	1,8237
	RGA	1,4921	0,0104	1,4903	0,0128		
	LHGA	**1,4766**	**0,0015**	**1,4651**	**0,0004**		
Housing	*GGA*	1,9622	0,0411	2,0063	0,0406	2,2358	2,2013
	RGA	1,9189	0,0013	1,9466	0,0189		
	LHGA	**1,8622**	**0,0014**	**1,8973**	**0,0169**		
Concrete Compressive Strength	*GGA*	5,3069	0,0461	5,3102	0,0475	6,1246	6,0237
	RGA	5,0706	0,0016	5,0824	0,0014		
	LHGA	**5,0680**	**0,0015**	**5,0609**	**0,0014**		
Concrete Slump Test	*GGA*	0,8027	0,0673	0,7284	0,0851	0,9876	0,9951
	RGA	0,6103	0,0423	0,6905	0,0852		
	LHGA	**1,8622**	**0,0306**	**0,6260**	**0,0502**		
Abalone (male)	*GGA*	5,3069	0,0011	2,0502	0,0012	2,1250	2,0872
	RGA	5,0706	0,0009	2,0285	0,0007		
	LHGA	**5,0680**	**0,0002**	**2,0222**	**0,0005**		
Parkinsons Telemonitoring	*GGA*	1,4921	0,0038	1,5013	0,0041	1,5324	1,5301
	RGA	1,4652	0,0007	1,4758	0,0007		
	LHGA	**1,4592**	**0,0007**	**1,4660**	**0,0007**		

However, the intelligent optimization algorithm does not greatly improve the criterion value for all examples. The reasons are two-fold. First, the variables in some examples are strongly correlated with each other. For the example of Heart Disease Index, the number of final variables selected by intelligent optimization algorithm, stepwise regression and complete model method is all 13. Secondly, the transformation uniformly applied may not suit all examples. For the example of Parkinsons Telemonitoring, the raw data exhibit certain periodicity, but power function, logarithmic function and exponential functions are used for the transformation without consideration of the periodicity. The model precision can be further improved by introducing other forms of functions that better suit the conditions of Parkinson's disease.

4.2 Performance comparison

To evaluate the performance of the improved genetic algorithm, on-line and off-line performance indicators proposed by De Jong are chosen [25]. These indicators are applied to large-sample-size examples Weather Ankara, Abalone (male) and Housing, and the performance is compared with that of GGA and RGA.

As shown in Fig. 1 ÷ 3, LHGA has a better on-line and off-line performance than RGA and GGA, indicating a better population performance and convergence. The randomness inherent in RGA may lead to instability. Under BIC, RGA is inferior to GGA for the examples of Parkinsons Telemonitoring and Weather Ankara in terms of on-line performance.

4.3 Comparison of prediction capacity

The prediction capacity is an important aspect of the regression model and is measured by mean square error (MSE). The smaller the MSE, the higher the prediction precision will be. The model obtained after each run was used for the prediction, and each example was run for 150 times. MSE was recorded for each run, and the average of 150 runs was taken.

As shown in Tab. 3, the average MSE with intelligent optimization algorithm under any rule is obviously

smaller than that of complete model method and stepwise regression. This indicates a higher prediction precision of the intelligent optimization algorithm. LHGA achieves the smallest MSE, except for the Housing dataset under AIC where the MSE is greater than that of RGA. Thus LHGA is superior to GGA and RGA in terms of prediction precision and stability.

5 Conclusion

By using LHS to modify the crossover operator in genetic algorithm, we propose an improved genetic algorithm that is a parallel implementation of variable selection, transformation and parameter estimation for the regression model. This method overcomes the defects of subjectivity, strong dependence on path and difficulty in processing massive high-dimensional data. Simulation experiments indicate that the improved genetic algorithm has high efficiency and stability when used to solve the regression model. This is a considerable improvement of the conventional regression techniques, especially in terms of the optimum seeking performance. Future studies will be oriented towards the effectiveness and progress through the application of LHGA to other statistical models.

Acknowledgments

Work is supported by the Education Department of Anhui Province Humanities and Social Sciences Key Planning Fund (Grant No. SK2015A563, SK2016A0971), the Education Department of Anhui Province Natural Science Key Research Projects(Grant No. KJ2016A742, KJ2013A259), and Anhui Province College Excellent Young Talents Support Plan Key Projects (2017), the Youth Scholars of the West Anhui University (Grant No. WXZR1510, WXSQ1432).

6 References

[1] Sandra, P.; Tommaso, M. Regression Model Selection Using Genetic Algorithms. // Recent Advances in Neural Networks, Fuzzy Systems & Evolutionary Computing. (2010), pp.19-27.

[2] Yang, C. Y.; Chuang, C. C.; Jeng, J. T.; Tao, C. W. Constructing the linear regression models for the symbolic interval-values data using PSO algorithm. // System Science and Engineering (ICSSE), 2011 International Conference on. IEEE. (2011), pp.177-181.

[3] Manouchehrian, A.; Sharifzadeh, M.; Hamidzadeh, M. R.; Nouri, T. Selection of regression models for predicting strength and deformability properties of rocks using GA. // International Journal of Mining Science and Technology. 23, 4(2013), pp. 495-501. DOI: 10.1016/j.ijmst.2013.07.006

[4] Liu, J. P.; Yu, J. X. Parameter estimation method of logistic regression models based on particle swarm optimization algorithm. // Computer Engineering and Applications. 45, 33(2009), pp. 42-44.

[5] Wu, N.; Huang, J.; Schmalz, B.; Fohrer, N. Modeling daily chlorophyll a dynamics in a German lowland river using artificial neural networks and multiple linear regression approaches. // Limnology. 15, 1(2014), pp. 47-56. DOI: 10.1007/s10201-013-0412-1

[6] Kialashaki, A.; Reisel, J. R. Modeling of the energy demand of the residential sector in the United States using regression models and artificial neural networks. // Applied Energy. 108 (2013), pp. 271-280. DOI: 10.1016/j.apenergy.2013.03.034

[7] Khashei, M.; Zeinal, H. A.; Bijari M. A novel hybrid classification model of artificial neural networks and multiple linear regression models. // Expert Systems with Applications. 39, 3(2012), pp. 2606-2620. DOI: 10.1016/j.eswa.2011.08.116

[8] Lv, C. L.; Chen, S. H.; Yang, Y. J. Simultaneous selection for variables and transformation in linear model. // Journal on Numerical Methods and Computer Applications. 4, 1(2005), pp. 26-35.

[9] Pham, D.; Karaboga, D. Intelligent optimisation techniques: genetic algorithms, tabu search, simulated annealing and neural networks. Springer Science & Business Media, 2012.

[10] Kang, F.; Han, S.; Salgado, R.; Li, J. System probabilistic stability analysis of soil slopes using Gaussian process regression with Latin hypercube sampling. // Computers and Geotechnics. 63 (2015), pp. 13-25. DOI: 10.1016/j.compgeo.2014.08.010

[11] Deutsch, J. L.; Deutsch, C. V. Latin hypercube sampling with multidimensional uniformity. // Journal of Statistical Planning and Inference. 142, 3(2012), pp. 763-772. DOI: 10.1016/j.jspi.2011.09.016

[12] Baudrit, C.; Dubois, D.; Perrot, N. Representing parametric probabilistic models tainted with imprecision. // Fuzzy sets and systems. 159, 15(2008), pp. 1913-1928. DOI: 10.1016/j.fss.2008.02.013

[13] Yu, H.; Chung, C. Y.; Wong, K. P.; Lee, H. Probabilistic load flow evaluation with hybrid latin hypercube sampling and cholesky decomposition. // Power Systems, IEEE Transactions on. 24, 2(2010), pp. 661-667. DOI: 10.1109/TPWRS.2009.2016589

[14] Crombecq, K.; Dhaene, T. Generating sequential space-filling designs using genetic algorithms and monte carlo methods.Simulated Evolution and Learning. Springer Berlin Heidelberg, 2010. DOI: 10.1007/978-3-642-17298-4_8

[15] Yu, H.; Chung C. Y.; Wong K. P.; Zhang J. A Probabilistic Load Flow Calculation Method with Latin Hypercube Sampling. // Automation of Electric Power Systems. 33, 21(2009), pp. 32-35.

[16] Vořechovský, M.; Novák, D. Correlation control insmall-sample Monte Carlo type simulations I: a simulated annealing approach. // Probabilistic Engineering Mechanics. 24, 3(2009), pp. 452-462. DOI: 10.1016/j.probengmech.2009.01.004

[17] Zhang L.; Zhang B. Good point set based genetic algorithm. // Chinese Journal of Computers. 24, 9(2001), pp. 917-922.

[18] Janssen, H. Monte-Carlo based uncertainty analysis: Sampling efficiency and sampling convergence. // Reliability Engineering & System Safety. 109 (2013), pp. 123-132. DOI: 10.1016/j.ress.2012.08.003

[19] Brown, R. W.; Cheng, Y. C. N.; Haacke, E. M.; Thompson, M. R.; Venkatesan, R. Magnetic resonance imaging: physical principles and sequence design. John Wiley & Sons, 2014. DOI: 10.1002/9781118633953

[20] Beck, J.; Hellekalek, P.; Hickernell, P. Random and quasi-random point sets. Springer Science & Business Media, 2012.

[21] Chen, M. H.; Zhou, B. D.; Ren, Z. Genetic algorithm based on random uniform design. // Applied Mathematics A Journal of Chinese Universities. 25, 3(2010), pp. 279-283.

[22] Cook R. D.; Weisberg S. Applied regression including computing and graphics. // John Wiley & Sons, 2009.

[23] http://funapp.cs.bilkent.edu.tr/DataSets//[EB/OL] (2015-8-5)

[24] http://archive.ics.uci.edu/ml/index.html (2015-8-5)

[25] Starkweather, T.; McDaniel, S.; Mathias, K.; Whitley, C.; Whitley, D. A comparison of genetic sequencing operators.

// In: Belew R, Booker L, eds. Proceedings of the 4th International Conference on Genetic Algorithms. Los Altos: Morgan Kaufmann Publishers. (1991), pp. 69-76.

Authors' addresses

Shi Minghua, Ph.D student, Associate Professor
1) Business school, University of Shanghai for Science and Technology, No. 334 Jungong Road, 200093, Shanghai, China
2) College of Finance and Mathematics & Financial Risk Intelligent Control and Prevention Institute, West Anhui University, No. 1 Yunluqiao West Road, 237012, Lu'an, China
E-mail: minghuashi@163.com

Xiao Qingxian, Professor
Business school, University of Shanghai for Science and Technology, No. 334 Jungong Road, 200093, Shanghai, China
E-mail: qxxiao@163.com

Zhou Benda, Professor
College of Finance and Mathematics& Financial Risk Intelligent Control and Prevention Institute, West Anhui University, No. 1 Yunluqiao West Road, 237012, Lu'an, China
E-mail: bendazhou@163.com

Yang Feng, Ph.D student, Associate Professor
1) Business school, University of Shanghai for Science and Technology, No. 334 Jungong Road, 200093, Shanghai, China
2) School of Information Technology, Henan University of Traditional Chinese Medicine, Longzihu University Park, No. 1 Zhengzhou, 450046, Henan China
E-mail: yangfeng1126@126.com