Regression model is a well-established tool in data analysis with applications in various fields. The selection of independent variables and the mathematical transformation of 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

**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 parameter estimation. Most of the existing algorithms can either deal with either variable selection or 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.

**LHS and its properties**

**2.1 LHS**

For a deterministic function \( Y = f(X) \), where \( Y \in R \)
$X \in \mathbb{R}^p$, $p$ statistically independent elements $X_1, X_2, \ldots, X_p$ of random vector $X$ 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 $[0, 1]^t$, $p_d(i) = \{(r_1, r_2, \ldots, r_t), i = 1, 2, \ldots, n\}$ where $r_t = 2\cos(2\pi nk)/p$, $1 \leq k \leq t$, and $p$ is the minimal prime larger than $2\pi/3$, or $r_t = e^i$, $1 \leq k \leq t$, 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 $(\pi_{ij}, \pi_{2j}, \ldots, \pi_{nj})$ one random arrangement of $n$ labels $(1, 2, \ldots, n)$ of coordinates in the $j$th dimension.
2. Suppose that the $t$ random arrangements are mutually independent, and the random matrix $\pi = (\pi_{ij})_{n^t}$ of order $n \times t$ is obtained; 3. Let

$$c_{ij} = \frac{\pi_{ij} + u_{ij}}{n}, i = 1, \ldots, n, j = 1, \ldots, t$$

(1)

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

### 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 L2 bias as more accurate measure. However, Hickernell [20] pointed out that bias and L2 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:

$$WD(P) = \left\{\left(\frac{4}{3}\right)^t + \frac{1}{n^t} \sum_{k=1}^n \prod_{j=1}^t \left[\frac{3}{2} - k_x - x_j \right] \left[1 - k_x - x_j \right]^{1/2}\right\}^t$$

where $x_i = (x_{i1}, x_{i2}, \ldots, x_{in})$; $n$ is the number of samples, $s$ is the dimensionality of the samples; $P = \{x_1, x_2, \ldots, x_n\}$ is the point set sampled over $C^t = \{0, 1\}^t$ cube.

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

$$E(WD(L_{n,s}))^2 = \left(\frac{4}{3}\right)^t + \frac{1}{n^t} \sum_{k=1}^n \prod_{j=1}^t \left[\frac{3}{2} - k_x - x_j \right] \left[1 - k_x - x_j \right]^{1/2}$$

**Prove:**

$$E(WD(L_{n,s}))^2 = \left(\frac{4}{3}\right)^t + \frac{1}{n^t} \sum_{k=1}^n \prod_{j=1}^t \left[\frac{3}{2} - k_x - x_j \right] \left[1 - k_x - x_j \right]^{1/2}$$

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

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

**Prove:**

$$E(WD(R_{n,s}))^2 = \frac{1}{n^t} \left[\frac{3}{2} - \left(\frac{4}{3}\right)^t\right]$$
\[
\frac{1}{n^2} \left( \frac{1}{3} \right)^y - \left( \frac{4}{3} \right)^y + \left( \frac{1 - 1}{n} \right) \left( \frac{3}{2} y - \frac{1}{2} y^2 + \frac{1}{3} y^3 \right) dy + \\
\frac{1}{n^2} \int_0^1 \left( \frac{3}{2} (1-y) + \frac{1}{2} (x-y)^2 + \frac{1}{3} (x-y)^3 \right) dy = \\
\frac{1}{n^2} \left( \frac{3}{2} \right)^y - \left( \frac{4}{3} \right)^y + \left( \frac{1 - 1}{n} \right) \left( \frac{3}{4} \frac{1}{6} + \frac{1}{3} \frac{1}{6} + \frac{1}{2} \frac{1}{6} \right) = \\
\frac{1}{n^2} \left[ \left( \frac{3}{2} \right)^y - \left( \frac{4}{3} \right)^y \right]
\]

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

**Prove:**

\[E(WD(R_{ds})^2) - E(WD(L_{ds})^2) = \left( \frac{1}{n^2} \left( \frac{4}{3} \right)^y - \left( \frac{4}{3} \left( 60y^2 - 60y \right) \right) \right) > 0\]

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 theoretically 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, \(p (p \leq m)\) independent variables are selected from \(m\) variable sets \(X = \{X_1, X_2, \ldots, X_m\}\) 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 \(X' = \{X_1', X_2', \ldots, X_p'\} \subseteq X\). 

\[\left\{ \hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p \right\}\]

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 Kullback-Leibler information. For the regression model, \(AIC = \log(S_y^2) + 2p/n\), where \(n\) is the number of samples; \(p\) is the number of independent variables in the regression equation; \(S_y^2\) 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, \(BIC = \log(S_y^2) + (p/n) \times \log n\).

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 \(T = \{0, 1, 2, 3\}\), \(E = \{-6, \ldots, -4, -3, \ldots, -1, 1, 2, 3, 4, 5, 6\}\), \(t_i \in T, e_i \in E\), then

\[t_i = 0 : T(X_i) = 0\]

\[t_i = 1 : T(X_i) = (X_i)^e_i\]

\[t_i = 2 : T(X_i) = (\log X_i)^e_i\]

\[t_i = 3 : T(X_i) = (e^{X_i})^e_i\]

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 \(A_i = (a_1^i, a_2^i, \ldots, a_k^i, a_{k+1}^i, \ldots, a_m^i)\) and \(A_j = (a_1^j, a_2^j, \ldots, a_k^j, a_{k+1}^j, \ldots, a_m^j)\), respectively. If \(i \leq 4\), then \(a_i^j \in T\); if \(4 < i \leq 16\), then \(a_i^j \in E\). Let \(H = \{(X_1, X_2, \ldots, X_n) | i \in J, x_i = \ast; i \notin J, x_i = a_i^j \}\), where \(J = \{i | a_i^j \neq a_i^j\}\) is the set composed of the loci of alleles \(A_1\) and \(A_2\), and let \(n(J) = t\). Then using the \(t\)-dimensional cubic composed of the loci of alleles, \(n\) points are selected for crossover by LSH.

Let the \(k\)-th chromosome in the \(n\)-th descendent be

\[b^{(k)} = (b_1^k, b_2^k, \ldots, b_k^k)\]

where

\[b_i^k = \begin{cases}
  a_i^j & i \notin J \\
  \frac{1}{t} (i - j) & i = j \in J, 1 \leq j \leq t, 1 \leq k \leq n, 1 \leq i \leq l
\end{cases}
\]

If \(i \leq 4\), then \(b_i^k \in T\) and \(\{a\}\) indicates that \(a\) is mapped from interval \([0,1]\) to the set \(\{0,1,2,3\}\) with squaring; if \(4 < i \leq 16\), then \(b_i^k \in E\), and \(\{a\}\) 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 $x_{\text{max}} = 6, x_{\text{min}} = -6$ and $y$ be the encoding after mapping. The mapping function is:

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

where $\text{sgn}(y_i)$ is the sign function of $y_i$. 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 descendnet 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 $\overline{f_{\text{run}}}$ and optimal variance $\rho_{\text{run}}$ in continuous runs were calculated.

**Definition [25]** Let

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

where $f_j(T)$ is the optimum obtained by genetic algorithm after $T$ iterations for the $j$-th time; $ho_{\text{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 $\overline{f_{\text{run}}}$, and the algorithm stability by $\rho_{\text{run}}$.

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>
<thead>
<tr>
<th>Problem</th>
<th>Alg</th>
<th>Maxval</th>
<th>Bestval</th>
<th>$\overline{f_{\text{run}}}$</th>
<th>$\rho_{\text{run}}$</th>
<th>Besttime</th>
<th>Complete Model</th>
<th>StepwiseRegression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Disease Index</td>
<td>GGA</td>
<td>17,0532</td>
<td>16,9947</td>
<td>17,0155</td>
<td>0.0108</td>
<td>58,3498</td>
<td>17,3402</td>
<td>17,2479</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>17,0452</td>
<td>16,9745</td>
<td>17,0057</td>
<td>0.0157</td>
<td>49,8660</td>
<td>17,2949</td>
<td>17,2479</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>17,0123</td>
<td>16,9711</td>
<td>17,0020</td>
<td>0.0082</td>
<td>43,6295</td>
<td>8,2210</td>
<td>8,2329</td>
</tr>
<tr>
<td>Weather Ankara</td>
<td>GGA</td>
<td>2,2932</td>
<td>2,0099</td>
<td>2,2126</td>
<td>0.2294</td>
<td>304,5379</td>
<td>2,3449</td>
<td>2,345</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>2,2105</td>
<td>2,0400</td>
<td>2,0809</td>
<td>0.1394</td>
<td>273,1118</td>
<td>2,3449</td>
<td>2,345</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>2,1817</td>
<td>1,9050</td>
<td>1,9656</td>
<td>0.1343</td>
<td>249,1413</td>
<td>2,3449</td>
<td>2,345</td>
</tr>
<tr>
<td>Housing</td>
<td>GGA</td>
<td>6,7865</td>
<td>6,4994</td>
<td>6,5909</td>
<td>0.0731</td>
<td>75,8487</td>
<td>8,2210</td>
<td>8,2329</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>6,5570</td>
<td>6,4658</td>
<td>6,5139</td>
<td>0.0311</td>
<td>90,1297</td>
<td>8,2210</td>
<td>8,2329</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>6,5573</td>
<td>6,4445</td>
<td>6,4904</td>
<td>0.0154</td>
<td>68,7977</td>
<td>8,2210</td>
<td>8,2329</td>
</tr>
<tr>
<td>Concrete Compressive</td>
<td>GGA</td>
<td>8,7678</td>
<td>8,5956</td>
<td>8,6069</td>
<td>0.0302</td>
<td>43,7629</td>
<td>10,080</td>
<td>10,0715</td>
</tr>
<tr>
<td>Strength</td>
<td>RGA</td>
<td>8,6159</td>
<td>8,5553</td>
<td>8,5667</td>
<td>0.0137</td>
<td>40,7920</td>
<td>10,080</td>
<td>10,0715</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>8,5967</td>
<td>8,3231</td>
<td>8,3334</td>
<td>0.0133</td>
<td>40,0896</td>
<td>10,080</td>
<td>10,0715</td>
</tr>
<tr>
<td>Concrete Slump Test</td>
<td>GGA</td>
<td>3,6208</td>
<td>2,8244</td>
<td>3,1901</td>
<td>0.1713</td>
<td>17,7092</td>
<td>4,5396</td>
<td>4,4868</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>3,0040</td>
<td>2,7887</td>
<td>2,9029</td>
<td>0.0361</td>
<td>18,4451</td>
<td>4,5396</td>
<td>4,4868</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>3,0045</td>
<td>2,6841</td>
<td>2,8068</td>
<td>0.0342</td>
<td>14,5150</td>
<td>4,5396</td>
<td>4,4868</td>
</tr>
<tr>
<td>Abalone (male)</td>
<td>GGA</td>
<td>4,7387</td>
<td>4,6401</td>
<td>4,6808</td>
<td>0.0223</td>
<td>353,1539</td>
<td>4,7640</td>
<td>4,7623</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>4,6593</td>
<td>4,6288</td>
<td>4,6651</td>
<td>0.0068</td>
<td>331,0419</td>
<td>4,7640</td>
<td>4,7623</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>4,6527</td>
<td>4,3632</td>
<td>4,4689</td>
<td>0.0047</td>
<td>120,8126</td>
<td>4,7640</td>
<td>4,7623</td>
</tr>
<tr>
<td>Parkinsons Telemonitoring</td>
<td>GGA</td>
<td>8,1447</td>
<td>7,9867</td>
<td>8,1849</td>
<td>0.0216</td>
<td>5820,3</td>
<td>8,3213</td>
<td>8,2943</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>8,0857</td>
<td>7,8491</td>
<td>8,0081</td>
<td>0.0077</td>
<td>6840,9</td>
<td>8,3213</td>
<td>8,2943</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>8,0660</td>
<td>7,7498</td>
<td>7,9542</td>
<td>0.0034</td>
<td>4980,7</td>
<td>8,3213</td>
<td>8,2943</td>
</tr>
</tbody>
</table>
Table 2 Comparisons between LHGA and the other 2 algorithms, BIC are considered

<table>
<thead>
<tr>
<th>Problem</th>
<th>Alg</th>
<th>Maxval</th>
<th>Bestval</th>
<th>$f_{run}$</th>
<th>$\rho_{run}$</th>
<th>Besttime</th>
<th>Complete Model</th>
<th>StepwiseRegression</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heart Disease Index</td>
<td>GGA</td>
<td>17,2041</td>
<td>17,1126</td>
<td>17,1629</td>
<td>0.0248</td>
<td>54,0396</td>
<td>17,5341</td>
<td>17,2959</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>17,2217</td>
<td>17,0917</td>
<td>17,1453</td>
<td>0.0303</td>
<td>51,8489</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>17,1896</td>
<td>16,9912</td>
<td>17,0357</td>
<td>0.0209</td>
<td>40,8431</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weather Ankara</td>
<td>GGA</td>
<td>2,2759</td>
<td>2,1175</td>
<td>2,2185</td>
<td>0.1445</td>
<td>289,6587</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>2,2436</td>
<td>2,0769</td>
<td>2,1074</td>
<td>0.1001</td>
<td>278,2713</td>
<td></td>
<td>2,3817</td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>2,1382</td>
<td>2,0694</td>
<td>2,0809</td>
<td>0.0090</td>
<td>239,4821</td>
<td></td>
<td>2,3778</td>
</tr>
<tr>
<td>Housing</td>
<td>GGA</td>
<td>6,8778</td>
<td>6,6027</td>
<td>6,6955</td>
<td>0.0586</td>
<td>86,4383</td>
<td>8,3359</td>
<td>8,3305</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>6,6827</td>
<td>6,5638</td>
<td>6,6312</td>
<td>0.0349</td>
<td>81,0427</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>6,6819</td>
<td>6,4687</td>
<td>6,5709</td>
<td>0.0178</td>
<td>70,2976</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concrete Compressive</td>
<td>GGA</td>
<td>8,8352</td>
<td>8,6108</td>
<td>8,6580</td>
<td>0.0324</td>
<td>42,9781</td>
<td>10,1195</td>
<td>10,1013</td>
</tr>
<tr>
<td>Strength</td>
<td>RGA</td>
<td>8,6555</td>
<td>8,5965</td>
<td>8,6137</td>
<td>0.0137</td>
<td>81,0427</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>8,6437</td>
<td>8,4031</td>
<td>8,4134</td>
<td>0.0124</td>
<td>38,4765</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Concrete Slump Test</td>
<td>GGA</td>
<td>3,7966</td>
<td>3,1101</td>
<td>3,4361</td>
<td>0.1584</td>
<td>18,6621</td>
<td>4,5466</td>
<td>4,5016</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>3,3099</td>
<td>3,0744</td>
<td>3,1823</td>
<td>0.0451</td>
<td>17,4561</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>3,3109</td>
<td>2,8585</td>
<td>2,9653</td>
<td>0.0410</td>
<td>14,0815</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Abalone (male)</td>
<td>GGA</td>
<td>4,7592</td>
<td>4,6702</td>
<td>4,7095</td>
<td>0.0176</td>
<td>322,1194</td>
<td>4,7949</td>
<td>4,7870</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>4,7012</td>
<td>4,6640</td>
<td>4,6764</td>
<td>0.0076</td>
<td>303,5799</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>4,6873</td>
<td>4,4632</td>
<td>4,4689</td>
<td>0.0055</td>
<td>112,9978</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Parkinsons Telemonitoring</td>
<td>GGA</td>
<td>8,1659</td>
<td>8,0777</td>
<td>8,1034</td>
<td>0.0224</td>
<td>6649,4</td>
<td>8,3435</td>
<td>8,3095</td>
</tr>
<tr>
<td></td>
<td>RGA</td>
<td>8,1037</td>
<td>8,0006</td>
<td>8,0037</td>
<td>0.0091</td>
<td>6531,1</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>LHGA</td>
<td>8,1034</td>
<td>7,8664</td>
<td>7,9776</td>
<td>0.0013</td>
<td>5240,3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 1(a) Weather Ankara performance result

Figure 1(b) Weather Ankara performance result

Figure 2(a) Abalone (male) performance result

Figure 2(b) Abalone (male) performance result
Proposed by De Jong are chosen [25]. These indicators be further improved by introducing other forms of consideration of the periodicity. The model precision can suit all examples. For the example of Parkinson’s stepwise regression and complete model method is all 13. examples are strongly correlated with each other. For the Parkinson’s Telemonitoring, the raw data exhibit certain periodicity, not greatly improve the criterion value for all examples. 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

Regression modeling based on improved genetic algorithm

Shi Minghua et al.


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