A Novel Spatial Interpolation Parallel Algorithm based on Genetic Expression Programming

Dongmei Zhang, Yang Li, Chengjun Li, Jianquan Bao

School of Computer Science, China University of Geosciences, Wuhan, 430074, China

jjiel@163.com, god_sun@163.com, cuglicj@gmail.com, 348322335@qq.com

Abstract

Because of the limited environmental monitoring stations, the spatial data we have obtained by monitoring is local, discrete and limited. Therefore, speculating the environmental pollution parameters without monitoring station by using known monitoring data for spatial interpolation has become a hot research topic. In our preliminary work, an approach combining Genetic Expression Programming (GEP) evolution modeling with Delaunay, is presented to achieve automatic spatial interpolation. However, as a kind of genetic, GEP has the problem of premature convergence. In this paper, a GEP combining parallel algorithm based on MPI is proposed. By adopting the master-slave coarseness parallel GEP algorithm for spatial interpolation and introducing the subgroups migrating policy based on ring topology, this proposed algorithm has realized the coarseness scalable parallel computing which can run in the processor within a certain size to raise the ability of jumping out of the local optimum thus can avoid the problem of premature convergence efficiently. The result of simulation experiment shows that, as to the normal dataset and abnormal dataset of SIC2004, the GEP parallel algorithm based on MPI has the same predictable accuracy as the serial algorithms do, and the execution time is much less than serial algorithm, which can better satisfy the real-time demand of monitoring the content of radioactive material.

Keywords: Spatial Interpolation, MPI, The Master-slave Coarseness Parallel Algorithm, GEP Parallel Algorithm

1. Introduction

The environmental monitoring data attaches great importance to the prevention and treatment of environmental pollution. However, because of the limited environmental monitoring stations, the spatial data we have obtained by monitoring is local, discrete and limited. Thus, how to speculate the environmental pollution parameters without monitoring station by fully using known monitoring data for spatial interpolation has become a hot research topic.

Nowadays, many spatial interpolation techniques have been used in this field. However, traditional methods, such as the inverse distance weighting method, Thiessen polygons method, polynomial regression method, Kriging method, etc [1], are based on the assumption that the space is smooth and continuous, so the interpolation accuracy of these methods is reliable for the areas with large monitoring stations density while the accuracy is much less reliable for the sparse sites or areas where the change of environmental pollution parameters is big. In recent years, many scholars have proposed improved algorithms and the interpolation models to solve the above problem. In 2003, Marquinez proposed the interpolation model which regards the five terrain variables (elevation, slope, aspect, the distance from the coast and the distance from the relative west) which can affect the precipitation as the impact factors to consider by using the multiple regression method and GIS technology [2]. Wong applied the neural network method, self-organizing method and fuzzy mathematics method to the interpolation model of precipitation, which had obtained the ideal interpolation result [3]. Daly had established PRISM (parameter-elevation regressions on independent slopes model), which is a kind of multiple regression method and is suitable for areas with large topographical relief [4]. In 2004, Lin proposed the IRBFN interpolation method, which has combined the standard radial basis function network with semi-variance function model and are applied to the spatial interpolation analysis of precipitation in Tanshui River in southern Taiwan. Compared with the standard radial basis function network and common Kriging method, the interpolation error of the estimated data obtained by this
model is smallest. Yaozhong Pan proposed a temperature spatial interpolation method SSI based on DEM and smart distance search, which can not only produce the grid temperature results with high precision and high spatial resolution, but also its interpolation results can reflect the zonal character of the temperature according to the change of longitude, latitude and altitude gradients objectively and in details[6]; Hong Yan applied the thin plate smoothing spline interpolation method of accurate DEM to the spatial simulation of climate in China, and it has reflected the basic characteristics of the climate space well in China[7]. In 2008, stepwise regression method and GIS technology were used by Suoquan Zhou for the analysis of the different time scales of the precipitation in the middle and upper reaches of the Yangtze River, which has effectively solved the problem that the spatial interpolation accuracy is not high in the complex terrain, and is suitable for different time scales[8]. For those areas where the monitoring stations are sparse and unreasonably distributed, although the above methods can improve the interpolation accuracy to some extent, they are so sensitive to the parameter design that the prior knowledge is needed to determine the corresponding parameters.

Based on the above methods, this paper combines the GEP with spatial correlation, and applies the method to the spatial interpolation model [11-12]. Although this method does not require prior knowledge in the modeling process, and it can improve the accuracy of interpolation, due to the strong search ability of GEP, it will produce many new structures with great changes during the genotype to phenotype conversion, and the variant, hybrid or other genetic operation of each gene location is likely to lead to great changes in the structure of tree. Therefore, in the search process, there are some good structure could not be preserved to continue to evolve thus reduce the convergence efficiency. In view of the above problems, there are two methods can be taken to improve the search efficiency of GEP: 1) Finding various new expression and operator to improve the disadvantages of GEP; 2) Committing to the design of parallel genetic algorithm to obtain better computing result and improve the performance of the algorithm. GEP has good characteristics of parallel processing and high parallel efficiency [13]. Based on this idea, this paper combines the GEP with parallel computing and applies to spatial interpolation model. The result of simulation experiment shows that the new algorithm can not only obtain better predictive results, but also can better satisfy the real-time demand.

This paper is organized as follows: In section 2, we introduce GEP and parallel computing, then we briefly describe the theory of applying the combination of the parallel computing and GEP to spatial interpolation model in section 3. The results and analysis of the simulation experiment and the validity of the new method proved by the simulation experiment are illustrated in section 4. The last section is devoted to conclusions and future work.

2. Principle of GEP and Parallel Computing

2.1 Gene Expression Programming

GEP is a new technology based on genetic algorithm (GA) and genetic programming (GP), which combines the fixed-length coding of GA with the non-linear entity expression during the decoding process of GP. Therefore, GEP is more efficient when solving complex problems. In GEP, Code is a kind of linear string with one or more genes at a fixed length, which is composed of the head and tail. The head contains variables and operators while the tail only contains variables. And the phenotype is the expression tree with different sizes and shapes. The procedure of GEP can be expressed as follows:

Step1 Produce the initialization groups;

Step2 Evaluate each individual's fitness. If it is consistent with the best individual, output the best individual and its optimum solution; otherwise turn to step3;

Step3 Execute the following genetic operators, then return to step2;

1) Retain the best individual; 2) Selection; 3) Reproduction; 4) Mutation; 5) Transform and insertion sequence elements; 6) Recombination.

Essentially, GEP has the characteristics of parallel processing. And GEP computing in parallel or distributed system can improve the speed of problem solving and the quality of solution.

2.2. Parallel Computing
Compared with traditional algorithm, the improvement of parallel genetic algorithm mainly involves two aspects: first, the single group of the serial genetic algorithm is divided into several sub-groups for divide and rule; second is to control and manage the information exchange among the sub-groups. Different Divide-and-Conquer algorithms can produce four kinds of parallel genetic algorithm structure and basic model: master-slave model, coarse-grained model, fine-grained model and mixed model parallel genetic algorithm.

1. Master-slave parallel genetic algorithm

In 1992, Abramson realized the master-slave parallel genetic algorithm in the parallel computer with shared memory, which contains only a single population distributed in the main processor, and the assessments of individual fitness are distributed to different processors. The structure of this parallel model is shown in Figure 1.

![Figure 1. The master-slave parallel genetic algorithm model](image)

The advantages of this algorithm are that it is easy to implement and the computing speed can be improved significantly when the communication overhead does not take the lead, while its disadvantages are that it is easy to result into the busy main process and idle child process or vice versa, which can cause the load imbalance, and low efficiency.

2. Coarse-grained parallel genetic algorithm

In coarse-grained parallel genetic algorithm, the whole population is divided into several sub-populations, which evolve independently most of the time, and only occasionally exchange the individuals during the operation, and this phenomenon is also known as migration. Compared with serial genetic algorithm, coarse-grained parallel genetic algorithm has a fundamental change in the operation method because the selection, hybridization and mutation operation of individuals are restricted within the sub-populations. The advantages of this method are that it is simple in principle, easy to implement, and its traffic is small and the experimental environmental requirement is low. Moreover, it is easy to use PVM or MPI to implement in workstation, PC group, or even a single machine.

3. Fine-grained parallel genetic algorithm

Such parallel genetic algorithm also has only one population, but the individual distribution in the population has a certain spatial structure for limiting the interaction among individuals. Usually, individuals are distributed evenly in a two-dimensional plane grid. The ideal situation is that each grid point can only deal with one individual, and the evolution of individual in the grid points is carried out by selecting father body for genetic operation in the field. Its advantages are that it is suitable for large-scale parallel machine with high parallelism degree while the disadvantages are the large communication capacity and overhead.

4. Mixed parallel genetic algorithm

Combining the above three methods, we can get the mixed parallel genetic algorithm. Its main advantage is that you can use the advantages of each method to improve the quality of solution and the convergence speed. The combination of two parallel genetic algorithms can form a hierarchical structure. The high level of most mixed parallel genetic algorithm adopts the coarse-grained parallel genetic algorithm of populations while the low level can adopt any kind of parallel model.

In this paper, a master-slave coarse-grained parallel GEP algorithm is adopted for spatial interpolation.

3. The Principle and Steps of Parallel GEP Algorithm
3.1. Design scheme

Based on the preliminary work, in this paper, the parallel computing is combined with GEP by using sub-groups migration strategy with ring topology for the spatial interpolation model. The basic idea of sub-groups migration strategy with ring topology can be expressed as follows: the population P is divided into M groups, and each group has N sub-populations (in the following description, $S_{X,0}$ is used to be the θ sub-population of group X, while $S_{0}$ is still the θ sub-population of P). In the beginning of the first round of migration, the whole $S_{1,1}$ migrate to $G_{2}$ to replace $S_{2,1}$, then $S_{2,1}$ migrate to $G_{3}$ to replace $S_{3,1}$, and so forth until $S_{M,1}$ migrate to $G_{1}$ and replace $S_{1,1}$. After migration, a corresponding interval is needed. In the second round of migration, the whole $S_{1,2}$ migrate to $G_{2}$ to replace $S_{2,2}$, then $S_{2,2}$ migrate to $G_{3}$ to replace $S_{3,2}$, and so forth until $S_{M,2}$ migrate to $G_{1}$ and replace $S_{1,2}$. Each round of migration can be done in accordance with the laws, and when it is the turn of $S_{X,N}$ in each group to migrate, the $S_{X,1}$ of each group needs to re-migrate in the next round of migration.

The programming design adopts the master-slave GEP parallel algorithm based on MPI, and its basic idea is that the master process receive individuals from each slave process and save them into the array in which the i individual expresses the individual received from the i slave process, then the master process conducts the individual migration among groups or in groups according to the migration intervals by using ring topology and sends the results of migration to corresponding slave process orderly, at last the slave process implements serial GEP and find the optimal individual for the master process.

3.2. Steps of Parallel GEP Algorithm

The basic steps are as follows:

3.2.1. The master process

Step 1. Reading data and sending the data to each slave process for population initialization.

Step 2. Judging whether the genetic algebra has reached the maximum, and if reached, then to make prediction for predictive data based on the current optimal model.

Step 3. Judging whether the current genetic algebra is the multiple of the migration intervals, and if not, the master process is blocked and waiting for messages from the slave process, then go to Step 2.

Step 4. The master process receives messages from each slave process, and then judges that whether the current genetic algebra is the multiple of the intervals of migration among groups, if it is, the master process will conduct the sub-population migration among groups according to the ring topology, otherwise, the master process will conduct the sub-population migration within the group and go to Step 2.

3.2.2 The slave process

Step 1. Receiving data from the master process, and using the received data for population initialization;

Step 2. Judging whether the genetic algebra is the multiple of the migration intervals, if not, then go to Step 4, otherwise, go to Step 3;

Step 3. Judging whether the current genetic algebra is the multiple of the migration intervals, if it is, each slave process sends their current optimal individual to the master process, otherwise, each slave process send its own individual to the master process randomly, and then each slave process receives messages from the master process;

Step 4. Each slave process implements serial GEP, adds one onto the genetic algebra and goes to Step 2.

4. Case Study

4.1. Data Source
The presented case study, from the Spatial Interpolation Comparison 2004 (SIC2004), is analyzed by the proposed method [14-15]. The datasets are split into normal dataset and emergency dataset. In the normal dataset, the training data set does not have the “outliers” as in the known data set, while the emergency dataset has the “outliers”. Both of the two datasets consist of 1008 sampling points, i.e., 200 locations used for training and 808 locations for testing. Figure 2 shows the distribution of the sampling points in the two training datasets. The sampling points are relatively flat with small variance in the normal dataset (left), while in the emergency dataset (right), the sampling points have two different points which are greatly different from the remaining points in size and color, which are called “hot points”.

![Figure 2. The distribution of the normal training dataset (left) and the emergency training dataset (right)](image)

4.2. Environment and Setting of Parallel Experiment

The simulation parallel experiment in this paper is realized in cluster environment based on MPI, where the cluster system is Dawn TC5000A high performance computing system. The configuration of the adopted cluster environment is as follows: there are three nodes, where each node has 24 CPU cores, and the frequency and memory of each CPU core are 2.6GHZ and 1G. The network configuration is Gigabit Ethernet, the operating system is UNIX and the parallel programming environment is mpich2-1.2.1.

The setting of relevant parameters of evolution modeling is shown in Table 1.

<table>
<thead>
<tr>
<th>Name of parameters</th>
<th>value</th>
<th>Name of parameters</th>
<th>value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Population size</td>
<td>150</td>
<td>IS and RIS transpositions probabilities</td>
<td>0.1</td>
</tr>
<tr>
<td>Probability of using constants in tail</td>
<td>0.1</td>
<td>One-point recombination probability</td>
<td>0.7</td>
</tr>
<tr>
<td>Number of genes</td>
<td>6</td>
<td>Two-point recombination probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Gene size</td>
<td>18</td>
<td>Gene recombination probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Function set</td>
<td>{+,*,/,sqrt,sin,cos,tan,exp}</td>
<td>Gene transposition probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Mutation probability</td>
<td>0.022</td>
<td>The max value of constant</td>
<td>9</td>
</tr>
<tr>
<td>Probability of using constants in head</td>
<td>0.01</td>
<td>The min value of constant</td>
<td>0</td>
</tr>
<tr>
<td>Probability of using variables in head</td>
<td>0.06</td>
<td>The max length of IS transpositions</td>
<td>3</td>
</tr>
<tr>
<td>The fitness threshold of normal dataset</td>
<td>0.950</td>
<td>The fitness threshold of abnormal dataset</td>
<td>0.989</td>
</tr>
</tbody>
</table>

In the parallel experiment, the total number of processors is fixed at 65, that are a master process and 64 sub-processes, and the setting of parameters of evolution modeling of each sub-process independently is shown in Table 1. The migration interval is set to 50, when the migration interval is reached, the optimal individual of sub-process or the random selected individual is sent to the master.
process, and then the order of the accepted 64 individuals are adjusted by master process and returned back to each sub-process to complete the individual migration.

In order to fully reflect the performance of parallel algorithm better, this paper makes evaluation from the quality of solution and the parallel speedup. Various types of evaluation are defined as follows:

Definition 1 (the evaluation index of solution quality) The evaluation of the solution quality mainly through the root mean square error (RMSE), mean absolute error (MAE), mean error (ME) and correlation coefficient ($\gamma$) four indexes for assessment. The calculation formula of each index is as follows:

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (t_i - \hat{t}_i)} \quad (1)
\]

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |t_i - \hat{t}_i| \quad (2)
\]

\[
ME = \frac{1}{N} \sum_{i=1}^{N} (t_i - \hat{t}_i) \quad (3)
\]

\[
\gamma = \frac{\sum_{i=1}^{N} (t_i - \bar{t}_i)(\hat{t}_i - \bar{\hat{t}})}{\sqrt{\sum_{i=1}^{N} (t_i - \bar{t}_i)^2 \cdot \sum_{i=1}^{N} (\hat{t}_i - \bar{\hat{t}})^2}} \quad (4)
\]

The smaller the values of the first three indexes are, the better the prediction result is. When the last index is higher, the modeling data is more similar to the predictive data.

Definition 2 (speedup $Sp$) the speedup of parallel is $Sp$, which is defined as the ratio of the average of serial execution time $\bar{T}_s$ and the average of parallel execution time $\bar{T}_p$. The calculation formula is as follows:

\[
Sp = \frac{\bar{T}_s}{\bar{T}_p} \quad (5)
\]

For each experiment, it runs independently 10 times and computes the average execution time, thus obtain the parallel speedup $Sp$.

4.3. Analysis of the Results

In this paper, we conduct the interpolation computing on the normal dataset and abnormal dataset of SIC2004 by using serial algorithm and parallel algorithm respectively.

4.3.1. Comparison of serial and parallel results on normal dataset

Figure 3 is prediction model of the best model obtained by the ten times modeling on normal dataset by using serial algorithm and parallel algorithm, where (a) is the real measured value, (b) is the prediction result by serial algorithm and (c) is the prediction result by parallel algorithm. From the figure we can see that both (b) and (c) are similar to (a), which can be concluded that both serial algorithm and parallel algorithm can predict the normal dataset.
Table 2 shows the predictive statistical properties and four error indexes of the best model obtained by ten times modeling on normal dataset by using serial algorithm and parallel algorithm respectively. From this table, it can be concluded that the prediction results on normal dataset of serial algorithm and parallel algorithm are very close, and both of them can obtain good prediction result which is close to real measured value.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev</th>
<th>MAE</th>
<th>ME</th>
<th>RMSE</th>
<th>Pearson's</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Value</td>
<td>57.00</td>
<td>180.00</td>
<td>98.02</td>
<td>98.80</td>
<td>20.02</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Serial Prediction</td>
<td>66.64</td>
<td>126.25</td>
<td>96.24</td>
<td>99.41</td>
<td>14.53</td>
<td>9.19</td>
<td>-1.78</td>
<td>13.19</td>
<td>0.758</td>
</tr>
<tr>
<td>Parallel Prediction</td>
<td>66.89</td>
<td>127.98</td>
<td>96.95</td>
<td>99.39</td>
<td>14.98</td>
<td>9.16</td>
<td>-1.07</td>
<td>13.07</td>
<td>0.759</td>
</tr>
</tbody>
</table>

4.3.2. Comparison of serial and parallel results on abnormal dataset

Figure 4 is prediction model of the best model obtained by the ten times modeling on abnormal dataset by using serial algorithm and parallel algorithm, where (a) is the real measured value, (b) is the prediction result by serial algorithm and (c) is the prediction result by parallel algorithm. From the figure we can see that both (b) and (c) are similar to (a) in normal region while there exists some errors in abnormal region, but both of them can predict the approximate location where the exception occurred, which also illustrates the effectiveness of this algorithm.
Table 3 shows the predictive statistical properties and four error indexes of the best model obtained by ten times modeling on abnormal dataset by using serial algorithm and parallel algorithm respectively. From this table, it can be concluded that except for the maximum and correlation coefficient, the prediction results on most predictive statistical properties and error indexes of serial and parallel algorithm are very close. For parallel algorithm, its predictive maximum is closer to real measured value, and its predictive correlation coefficient is slightly better than the prediction of serial algorithm.

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>Max.</th>
<th>Mean</th>
<th>Median</th>
<th>Std. Dev</th>
<th>MAE</th>
<th>ME</th>
<th>RMSE</th>
<th>Pearson’s R</th>
</tr>
</thead>
<tbody>
<tr>
<td>Real Value</td>
<td>57.00</td>
<td>1528.00</td>
<td>105.40</td>
<td>98.95</td>
<td>83.71</td>
<td>-</td>
<td>-</td>
<td>64.79</td>
<td>0.678</td>
</tr>
<tr>
<td>Serial Prediction</td>
<td>64.94</td>
<td>698.44</td>
<td>102.83</td>
<td>100.50</td>
<td>36.48</td>
<td>16.50</td>
<td>-2.59</td>
<td>64.79</td>
<td>0.699</td>
</tr>
<tr>
<td>Parallel Prediction</td>
<td>69.46</td>
<td>986.33</td>
<td>101.72</td>
<td>95.99</td>
<td>47.34</td>
<td>16.27</td>
<td>-3.70</td>
<td>60.99</td>
<td></td>
</tr>
</tbody>
</table>

4.3.3 Speedup

Speedup is a main index of measuring the parallel performance of the system. When computing the speedup, the normal and abnormal datasets are run independently 10 times. The execution time will be recorded respectively each time when the fitness of serial and parallel program have reached the given threshold, and then find the average, then calculate the speedup $Sp$. The speedup of parallel algorithm relative to the serial algorithm is shown in Table 4.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Average Execution Time (Second)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>Normal Dataset</td>
<td>Serial Algorithm 2851.46</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Parallel Algorithm 470.86</td>
<td></td>
</tr>
<tr>
<td>Abnormal</td>
<td>Serial Algorithm 10120.41</td>
<td>5</td>
</tr>
</tbody>
</table>
It is shown from above table that for normal and abnormal dataset, the needed execution time to reach the given fitness threshold by using parallel algorithm is much less than the needed time by using serial algorithm, that is the parallel algorithm can meet the real-time requirement better.

5. Conclusions

In our previous work, we proposed a new approach Gene expression programming for spatial interpolation. In addition, combining the Delaunay algorithm with GEP can make full use of the local spatial structure information of variable, which can enhance the descriptive ability of information spaces’ distribution and weaken the noise impact of the predict data The results indicate that the proposed approach can identify the center position of the extreme values. Based on these researches, in this paper, we combine GEP with parallel algorithm by using potential parallel ability of GEP to reduce the execution time of searching for optimal solutions. This proposed algorithm has realized the coarseness scalable parallel computing which can run in the processor within a certain size to raise the ability of jumping out of the local optimum thus can avoid the problem of premature convergence efficiently. The result of simulation experiment shows that, as to the normal dataset and abnormal dataset of SIC2004, the GEP parallel algorithm based on MPI has the same predictable accuracy as serial algorithm, and its needed execution time is much less than serial algorithm, which can better satisfy the real-time demand of monitoring the content of radioactive material.

Acknowledgement

This work is support by the Natural Science Foundation of China (Grant No. 40972206), the National High-Tech Research and Development Plan of China (Grant Nos.863-317-01-04-99, 2008AA12A201), the Special Fund for Basic Scientific Research of Central Colleges, China University of Geosciences (Wuhan) (Grant Nos. 1323520909).

References