ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 164

International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)

Research on Image Clustering Based on Evolutionary Programming Algorithm

She Liu and Shuying Yang

[1]Department of computer and Communication Engineering, Tianjin University of Technology, Tianjin 300384 China

[2]Key laboratory of Intelligence Computing and Novel Software Technology, Tianjin University of Technology, Tianjin 300384 China

***Abstract*—Evolutionary programming algorithm is applied to the problem of image clustering. The solution of the problem is encoded by the symbol coding, and swarm intelligent model is used to search the solution of the problem. In evolutionary programming algorithm, the global search ability is effectively improved by mutation operator and selection operator. The excellent diversity of solutions is guaranteed by using Gaussian mutation operator, and the complexity of the evolutionary operation is reduced. The simulation experiments show that the proposed algorithm for image clustering is effective and correct.**

Keywords—evolutionary programming; image clustering; swarm intelligence

I. INTRODUCTION

Image clustering [1] refers to classifying many images into different categories, and the ones with similar content into one category. It occupies an important position in the field of pattern recognition and is mainly used for data compression, data mining, image retrieval and so on. Image clustering involves feature extraction and recognition algorithms, and different algorithms have different speed and recognition accuracy. Genetic algorithm, K - means algorithm and particle swarm optimization algorithm are commonly used for this purpose. Genetic algorithm [2 - 3] has poor performance in the local search process, relatively slow search speed and is prone to premature phenomenon. K - means algorithm [4] is sensitive to the initial clustering center and is likely to fall into local optimum. Particle swarm optimization algorithm [5] has the disadvantages of slow convergence, low accuracy and local minima problem. In clustering, the above-mentioned algorithms will inevitably do some repeated searches. If the number of samples is very large, the time consumption of the above-mentioned algorithms will be a serious drawback. [6].

Evolutionary Programming [7] (Evolutionary Programming, EP) is a machine evolutionary model proposed to solve prediction problems. As a random search method, it simulates natural evolutionary process and belongs to a branch of evolutionary computation. This algorithm introduces normal distribution technology into mutation operation and it is widely used in evolutionary computation [8]. Evolutionary programming can be applied to solving combinatorial optimization problems and complex nonlinear optimization problems, requiring only that the problems are computable and has been applied to many fields such as artificial intelligence and robot intelligent control [9 - 10].

Compared with genetic algorithm and evolution strategy [11], evolutionary programming mainly focuses on the evolution of species in the process of simulating biological evolution, never using individual crossover operators and recombination operators. The selection operator in evolutionary programming focuses on competitive selection among individuals in a group, but when the number of competing individuals is large, it is similar to the selection process in evolutionary strategy. Evolutionary programming directly takes the feasible solution of the problem as the coding of the individual, it isn't necessary to code for the individual, which is more convenient for application. In this study, evolutionary programming optimization algorithm is applied to image clustering problem, and the performance of the algorithm is verified by simulation.

II. BASIC PRINCIPLES OF EVOLUTIONARY PROGRAMMING

(1) Population initialization (randomly distributed individuals)

(2) Variation (updating individuals)

In evolutionary programming, Gaussian mutation operator is used to achieve individual mutation within the population in order to maintain the rich diversity. The standard deviation of Gaussian mutation operator changes according to individual fitness. As a result, for individuals with poor fitness, the variation range is large, and the search range is expanded. For individuals with high fitness, the variation range is small, the search range is very near to the local.

(3) Fitness calculation (evaluation of individual)

(4) Selection (group renewal).

Compared with other evolutionary computation[12], evolutionary programming algorithm does not use crossover of genetic algorithm or gene recombination operator of evolutionary strategy algorithm, but mutation operator operation reflects interaction between individuals.

III. IMPLEMENTATION OF CLUSTERING ALGORITHM

The steps of clustering algorithm based on evolutionary programming are as follows:

(1) Set up parameters: Initialize the size of the number of individuals $Size_{pop}$, the maximum iteration number $Iter_{max}$, the number of clustering centers Num_{center} and the number of individuals used for comparison in tournament competition q.

(2) Obtain the number and features of all samples to be clustered.

(3) Initialize the population, randomly assigning the values of each individual gene locus.

(4) Calculate the fitness of each individual.

(5) Generate $Size_{pop}$ new individuals through Gaussian mutation operator: for every gene locus of each individual, using the mutation operator, assigning a number between 1~-$Num_{center}$to the locus and thus generating the offspring.

(6) Calculate the individual fitness value of the $Size_{pop}$ offspring.

(7)Let the parent and child individuals ($2\times Size_{pop}$ in total) make the new group totalpop , and carry out selection operator operation on the new group: Randomly select q individuals from the totalpop, let each individual compare the fitness with the q individuals one by one, take the number of q individuals whose fitness are higher than the current individual fitness as the score of the current individual, and finally select the $Size_{pop}$ individuals with the highest score as the next generation parent.

(8) Record the optimal solution in the population: If the fitness of the optimal offspring individual is lower than the fitness of the total optimal individual (the closer the distance between them is, the smaller the fitness), the total best individual will be replaced by the current best individual.

(9) Judge whether the stop condition is met: the total iteration number is reached. If yes, the optimal solution is output, i.e. the category number of each sample is output and exits; If no, jump to step (4).

IV. Image Clustering Design Based on Differential Evolution Algorithm

A. Feature Extraction

The method for extracting featuresin this study is to delimit a rectangular frame at the outer edge of each number, and to segment the length and width of the rectangular frame by N× M, and each word after segmentation is called N×M template.Then calculate the proportion of the number of black pixels in each small square in this square as the sample features. The sample is divided into 7× 5 small squares, and the proportion of black pixels in the total square area is counted in each small square to obtain the features. the values of n and m can be appropriately modified according to the needs of practical problems. The larger the value, the more the sample features, the higher the clustering accuracy, and the algorithm convergence time will be correspondingly prolonged. The smaller the value is, the faster the algorithm converges, but the clustering accuracy will decrease.

B. Coding of Solutions

In the evolution programming, symbol coding is adopted, the bit string length is L, and the search space is an L - dimensional space, each element of which is an L - dimensional vector. In the algorithm, each chromosome X that makes up the evolutionary population can be represented directly by this L - dimensional vector.

Figure 1 shows the samples to be clustered, in which each individual contains one classification scheme. L take the twelve bit, the gene represents the class number (1 - 4) to which the sample belongs, the sequence number of the gene bit represents the number of the sample, the sequence number of the gene bit is fixed, that is to say, the position of a sample in the chromosome is fixed, and the class to which each sample belongs is changing at any time.If the gene bit is n, it corresponds to the sample number n, and the gene value pointed to by the nth gene position represents the belonging class number of the nth sample. Each solution contains one classification scheme, and the initial chromosome code of an individual is (1, 3, 4, 1, 2, 4, 2, 3, 1, 3, 2, 1), which means that samples 1, 4, 9, and twelve are classified into category 1.Samples 5, 7 and 11 were classified into category 2;Samples 2, 8 and 10 were classified into category 3;The 3rd and 6th samples belong to the 4th category, and are still in the hypothetical classification case, not the optimal solution, as shown in Table 1.

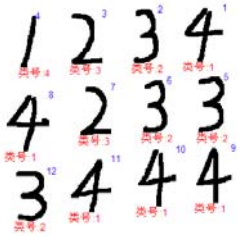

FIGURE I. EFFECT DIAGRAM OF HANDWRITTEN NUMERAL CLUSTERING

table i. OPTIMAL SOLUTION CODING FOUND BY EVOLUTIONARY PROGRAMMING ALGORITHM

Sample value	(4)	(3)	(2)	(1)	(3)	(3)	(2)	(4)	(4)	(4)	(4)	(3)
Gene value	1	2	3	4	2	2	3	1	1	1	1	2
Gene position	1	2	3	4	5	6	7	8	9	10	11	12
Sample number	1	2	3	4	5	6	7	8	9	10	11	12

C. Set the Fitness Function

The fitness function represents the degree of superiority and inferiority of each individual and calculates its fitness value *fitness* for each chromosome in the initial population. Its steps are as follows:

(1) Obtain the total number of cluster categories Num_{center} through manual intervention.

(2) Find out the sample with the same class number in the chromosome, $X^{(i)}$ represents the sample in class i.

(3) Count the number of samples n of each class, n_i is the number of samples in class i, the total number of samples is $popSize = \sum_{i}^{centerNum} n_i$.

(4) Calculating the center C of the same class.

$$C_i = \frac{1}{n_i}\sum_{k=1}^{n_i} X_k^{(i)}, \ (i = 1,2,\cdots,centerNum) \quad (1)$$

(5) Calculating the distance from each sample to the corresponding cluster center in the same class, and summing the distances

$$D_i = \sum_{j=1}^{n_i} \left\| X_j^{(i)} - C_i \right\|^2 \quad (2)$$

(6) Sum different kinds of D_i and take the *m_pop(i).fitness* as the most fitness function. The smaller $m\_pop(i).fitness$ is, the smaller error of this classification method is, that is, the smaller the fitness.

$$\mathrm{m}\_pop(i).fitness = \sum_{i=1}^{centerNum}\sum_{j=1}^{n_i}\left\| X_j^{(i)} - C_i \right\|^2 = \sum_{i=1}^{centerNum} D_i \quad (3)$$

D. Mutation Operator

Variation operation is the most important operation and the only search method in evolutionary programming algorithm, which is unique to evolutionary programming.Let X be the target variable σ of chromosome individual solution, and let it be the standard deviation of Gaussian variation. Chromosome consists of target variable x and standard deviation σ, as $(X,\sigma)=((x_1,x_2,\ldots,x_L),\sigma)$.

The form of individual variation is as follows: $X(t+1) = X(t) + N(0,s)$

The relationship between x and variance σ is

$$\begin{cases} s(t+1) = \sqrt{bF(X(t)) + g} \\ x_i(t+1) = x_i(t) + N(0, s(t+1)) \end{cases} \quad (4)$$

In formula (4): $F(X(t))$ represents the fitness value of the current individual, the closer the fitness value is to the target solution individual, the smaller the fitness value is.; $N(0,s)$ is a Gaussian random variable with probability density $p(s) = \frac{1}{\sqrt{2p}}\exp(-\frac{s^2}{2})$; The coefficients β and γ are undetermined parameters, the values of β and γ are generally set to 1 and 0; Each component of vector x can achieve different mutation effects through formula(4).

E. Selection Operator

The specific process of tournament selection is:

(1) Combining the population $P(t)$ composed of parents of $Size_{pop}$ and the population $P(t)$ composed of children of size generated by the population professionals after a mutation operation to form a collection $P(t) \cup P'(t)$ containing $2\times Size_{pop}$ individuals, as ***I.***

(2) Each individual $x_i \in \boldsymbol{I}$. Randomly selected q individuals from ***I***, and compared the fitness function values ($j \in (1,2,\ldots,q)$) of q individuals with the fitness function values of x_i, calculated that the number of individuals whose fitness function values are worse than x_i's fitness is w_i, and took wireless as x_i's score, where $w_i \in$ (0,1,..., q).

(3) After all the $2\times Size_{pop}$ individuals have been compared, they are sorted w_i according to the scores of each individual, and the $Size_{pop}$ individuals with the highest scores are selected as the next generation population.

In the process of evolution, relatively good individuals in each generation of population are given a larger score and can be retained in the next generation of population.

V. Experimental Results

Under the environment of MATLAB 2008 , the clustering problem of handwritten numerals based on evolutionary programming is implemented, and the images to be clustered as shown in fig. 1on the computer of Inter(R) Core(TM) i5 CPU , the experimental results are shown in figure 1

The optimal solution codes found by the evolutionary programming are shown in Table 1. Same handwritten numerals are classified into one category, assigned to the same category number, and all are correct.

VI. Conclusion

This study analyzes the principle and mechanism of evolutionary programming and applies it to image clustering. Simulation shows that the image clustering method based on evolutionary programming algorithm has good feasibility and accuracy.

References

[1] Zhang Shuo She, Xu Zongben, Liang Yi;Global Annealing Genetic Algorithm and Its Necessary and Sufficient Conditions for Convergence [J];China Science E Series;2002, 1997

[2] Zhang Zhonghua, Yang Shuying.Image clustering design based on genetic algorithm [J].Measurement and control technology in 2010, 29 (2): 4 - 46.

[3] Guo Qingrui, Xu Jianlong, Sun Shusen, et al.Color image clustering segmentation algorithm based on color barycenter and K - means [J].Journal of Zhejiang University of Technology, 2010, 27 (4): 580 - 584.

[4] Song Jie, Dong Yongfeng, Hou Xiangdan, etc.Improved particle swarm optimization algorithm [J].Journal of Hebei University of Technology, 2008, 37 (4): 55 - 59.

[5] Xie Jinxing.A brief summary of evolutionary computation [J].Control and decision - making. 1997, 12 (1): 1 - 7.

[6] Li Zhenying, Yao Ming.Cited in this paper [.IEEE transactions in evolutionary computing.2004, 8 (2): 1 - 13.

[7] Mr Yansong.Cited in this paper [.Computer physical communication.2002, 147 (8) : 729 - 732.

[8] Shang Yunwei, Qiu Li Huang.A Fast Evolutionary Programming Method for Solving Numerical Optimization Problems [J].Journal of System Simulation. 2004,16 (6): 1190 - 1192

[9] Wang Liping, Dong Jianghui.Convergence analysis of evolutionary programming algorithm with a new operator [J].Science, Technology and Engineering. 2009, 9 (6): 1428 - 1431

[10] Gao Yongchao;Li Qiqiang;;Annealing Evolutionary Programming Algorithm and Its Convergence [J];Journal of System Simulation;Gao Yongchao, 03, 2006;

[11] Wang Xiangjun, Ji Dou, Zhang Min;A Multi - group Competitive Evolutionary Programming Algorithm [J];Journal of electronics;11 issues in 2004

[12] Gao Wei.A Comparative Study of Genetic Algorithm and Evolutionary Programming [J].Communications and computers. 2005, 2 (8): 10 - 15.