Article ID: 1000-8152(2004)06-0885-05

Hypothesis-test based genetic algorithm for stochastic optimization problems

ZHANG Liang, WANG Ling, ZHENG Da-zhong
(Department of Automation, Tsinghua University, Beijing 100084, China)

Abstract: To effectively solve the stochastic optimization problems with non-deterministic and multi-modal properties, a class of hypothesis-test based genetic algorithm is proposed. The algorithm performs reasonable estimation by multiple evaluations, searches the design space effectively via genetic operators, and enhances the searching ability and population diversity by hypothesis test to overcome premature convergence. Based on typical stochastic functional and combinatorial optimization problems, the effects of hypothesis test, performance estimation number and magnitude of noise on the performance of the approach are studied, and the effectiveness and robustness of the proposed approach are demonstrated.

Key words: genetic algorithm(GA); stochastic optimization; hypothesis test
CLC number: TP18 **Document code**: A

随机优化问题基于假设检验的遗传算法

张 亮, 王 凌, 郑大钟
(清华大学 自动化系,北京 100084)

摘要: 为了有效解决具有不确定性和多极小性的随机优化问题,提出了一类基于假设检验的遗传算法.该方法通过多次评价来进行解性能的合理估计,利用遗传操作来进行解空间的有效搜索,采用假设检验来增加种群的多样性和算法的探索能力,从而避免遗传算法的早熟收敛.基于典型的随机函数优化和组合优化问题,仿真研究了假设检验、性能估计次数、噪声幅度对算法性能的影响,验证了所提方法的有效性和鲁棒性.

关键词: 遗传算法; 随机优化; 假设检验

1 Introduction

Generally speaking, stochastic optimization problems can be described as follows:

$$\min_{\theta} J(\theta) = \min_{\theta} \mathrm{E}[L(\theta,\xi)], \tag{1}$$

where θ is decision solution, ξ represents the stochastic effects of the system, $L(\theta,\xi)$ and $J(\theta)$ are the sample performance and its expected value is θ.

Stochastic optimization problems are often structureless, non-deterministic, and short of objective function explicitly known, so that the performance evaluation is done only by simulation. Meanwhile, the search space is often huge and there are many local optimum so that it is very hard to achieve global optimum. Currently, the study on stochastic optimization or simulation optimization has been a hot topic in the international academic fields[1], especially the research on designing effective and robust algorithms. In recent years, intelligent optimization algorithms have gained wide attention in both theoretical and engineering fields[2], but much for deterministic problems. Genetic algorithm (GA) is a kind of optimization algorithm based on the principles inspired by natural evolution phenomena, which has been widely and successfully applied in many fields[3], but it is very prone to be premature. Aiming at the non-deterministic property of stochastic optimization problem, Bayer[4] proposed some theoretical issues to study evolutionary algorithms in noisy environment. From the statistics viewpoint, a class of general GA based on hypothesis test is proposed in this paper, which applies hypothesis test based on mean value comparison to enhance population diversity so as to avoid premature convergence and improve the effectiveness of search space exploration via genetic operators. Based on stochastic functional and combinatorial optimization problems, the effects of hypothesis test, performance estimation number and noise magnitude on performance are studied, and the effectiveness and robustness of the proposed algorithm are demonstrated.

2 Hypothesis test

Hypothesis test is an important statistical method that is used to make test for predefined hypothesis using experi-

Received date: 2003-06-23; Revised date: 2004-03-30.
Foundation item: Supported by the National Natual Science Foundation of China (60204008, 60374060); 973 Program(2002CB312200).

ment data[5]. To make hypothesis test for different solutions when optimizing the stochastic problem, it often needs multiple independent simulations to provide suitable performance estimation for decision solutions. If n_i independent simulations are carried out for solution θ_i, then its unbiased estimated mean value $\bar{J}_i$ and variance s_i^2 can be calculated as follows:

$$\bar{J}_i = \bar{J}(\theta_i) = \sum_{j=1}^{n_i} L(\theta_i, \xi)/n_i, \tag{2}$$

$$s_i^2 = \sum_{j=1}^{n_i} [L(\theta_i, \xi) - \bar{J}_i]^2/(n_i - 1). \tag{3}$$

Considering two different solutions θ_1 and θ_2, whose estimated performances $\hat{J}(\theta_1)$ and $\hat{J}(\theta_2)$ are two independent random variables. According to the law of large number and central limit theorem, the estimation $\hat{J}(\theta_i)$ subjects to $N(\bar{J}_i, s_i^2/n_i)$ when n_i approaches to ∞. Suppose $\hat{J}(\theta_1) \sim N(\mu_1, \sigma_1^2)$ and $\hat{J}(\theta_2) \sim N(\mu_2, \sigma_2^2)$, where the unbiased estimation values of μ_1, μ_2 and s_1^2, s_2^2 are given by Eq.(2) and Eq.(3), and let the null hypothesis H_0 be "$\mu_1 = \mu_2$" and the alternative hypothesis H_1 be "$\mu_1 \neq \mu_2$".

If σ_1^2 and σ_2^2 are known, then the critical region of H_0 is described as follows:

$$|\bar{J}_1 - \bar{J}_2| \geqslant z_{\alpha/2}\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \tau, \tag{4}$$

where α is the evidence level with the meaning that $\phi(z_{\alpha/2}) = 1 - \alpha/2$.

If σ_1^2 and σ_2^2 are unknown and n_1, n_2 are large enough (say 50), then the critical region of H_0 can be simplified as follows:

$$|\bar{J}_1 - \bar{J}_2| \geqslant z_{\alpha/2}\sqrt{s_1^2/n_1 + s_2^2/n_2} = \tau, \tag{5}$$

where $s_1^2 = \sum_{j=1}^{n_1} [L(\theta_1, \xi) - \bar{J}_1]^2/(n_1 - 1)$ and $s_2^2 = \sum_{j=1}^{n_2} [L(\theta_2, \xi) - \bar{J}_2]^2/(n_2 - 1)$ are the unbiased estimation of σ_1^2 and σ_2^2 respectively.

If $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and σ^2 is unknown, then the critical region of H_0 is described as follows:

$$|\bar{J}_1 - \bar{J}_2| \geqslant t_{\alpha/2}(n_1 + n_2 - 2) \cdot \sqrt{[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2]/(n_1 + n_2 - 2)} \cdot \sqrt{(n_1 + n_2)/(n_1 n_2)} = \tau. \tag{6}$$

Thus, if $|\bar{J}(\theta_1) - \bar{J}(\theta_2)| < \tau$, i.e. the null hypothesis holds, then it can be regarded that the performances of these solutions are not significantly different from each other in statistical sense; otherwise they are indeed different. Furthermore, for stochastic minimization problem it is assumed that θ_2 is better than θ_1 if $\bar{J}(\theta_1) - \bar{J}(\theta_2) \geqslant \tau$, while θ_1 is better than θ_2 if $\bar{J}(\theta_1) - \bar{J}(\theta_2) \leqslant -\tau$. In addition, for specific problem it often supposes that the theoretical performance variances of all solutions are the same[5], so the hypothesis test can be made according to Eq.(6). For multi-modal stochastic optimization problems, a comparison under pure hypothesis test can often be trapped into local optima, so that it motivates us to combine the hypothesis test with the effective search ability of GA.

3 Hypothesis-test based genetic algorithm

Based on the idea of hypothesis test, after all new solutions are generated by genetic operators for stochastic optimization problems, they will be firstly ordered according to their estimated mean performances from the best to the worst, and the first solution will be put into the next population. Then, one by one from the second solution to the last solution, the current solution is compared with its nearest former solution not discarded. If there is no significant difference between their performances (say null hypothesis holds), then the current solution will be discarded to avoid repeated search; otherwise reserve the solution and put it into next population. After finishing such comparison-based hypothesis test, all those discarded solutions will be replaced by new solutions randomly generated, which will be put into next population to enhance diversity to some extent. Thus, a hypothesis-test based GA (HTGA) for stochastic optimization problem is proposed as follows.

Step 0 Given parameters such as population size P_s, mutation probability P_m, let $k = 0$.

Step 1 Randomly generate initial population $P(0)$, and estimate the performance $\bar{J}_i$ and variance s_i^2 of each solution with multiple (n times) independent simulations.

Step 2 Let the best solution of $P(k)$ be θ^* with the estimated performance $\bar{J}^*$ and variance s_*^2. If the stopping condition is satisfied, then output the best solution and its performances, otherwise go on to the steps below .

Step 3 Order all the solutions of $P(k)$ by $\bar{J}_1 \leqslant \bar{J}_2 \leqslant \cdots \leqslant \bar{J}_{P_s}$.

Step 4 Repeat $P_s/2$ times genetic operators (including selection, crossover, and mutation) for $P(k)$ to get a temporary population $P'(k)$, and estimate the performance $\bar{J}_i^t$ and variance $s_i^{t^2}$ of every new solution by multiple independent simulations.

Step 5 Order all the solutions of $P^t(k)$ by $\bar{J}_1^t \leqslant \bar{J}_2^t \leqslant \cdots \leqslant \bar{J}_{P_s}^t$, and denote the resulted solutions by $\theta_1^t, \theta_2^t, \cdots, \theta_{P_s}^t$ respectively. Let $m = 1, j = 2$, and put θ_1^t into $P(k+1)$ and denote it as θ_m.

Step 6 Perform the hypothesis test for θ_j^t with θ_m which is in $P(k+1)$. If the null hypothesis holds, i.e. Eq.(6) does not hold, then θ_j^t is discarded from $P^t(k)$; Otherwise, θ_j^t is put into $P(k+1)$ and denoted as θ_{m+1}, and let $m = m+1$.

Step 7 If $j < P_s$ then let $j = j+1$ and go to Step 6; Otherwise randomly generate $p_s - m$ new solutions and put them into $P(k+1)$, and replace the worst solution of $P(k+1)$ by the best solution of $P(k)$, let $k = k+1$, then go to Step 2.

It can be seen from the above procedure that, firstly the algorithm inherits the fundamental framework and operators of GA to keep the generality and effective optimization ability. Secondly, for stochastic optimization problems the hypothesis test based on statistical performance can enhance the population diversity, avoid the repeated search, and reasonably deal with the random factor to some extent. Besides, the elitist strategy in Step 7 guarantees the reservation of best solution found so far. Next we will demonstrate the effectiveness and robustness of HTGA by simulation based on stochastic functional and flow shop problems.

4 Numerical simulation and analysis

4.1 Study on stochastic functional optimization

Considering the following two-dimensional random Rosenbrock function:

$$L(x_1, x_2, \xi) = 100(x_1^2 - x_2)^2 + (1-x_1)^2 + \eta \cdot \xi, \quad |x_1, x_2| \leqslant 2.048, \tag{7}$$

where η denotes noise magnitude, ξ is random noise subjected to $N(0,1)$. Theoretically, the optimal solution of $J(x_1, x_2) = \mathrm{E}[L(x_1, x_2, \xi)]$ is $(1,1)$ with the best performance 0.

Firstly, apply the simple elitist GA without hypothesis test (random selection, arithmetic crossover and Gaussian mutation)[2], and performance estimation is based on only one simulation ($n = 1$). Let $\eta = 0.01, P_s = 30, p_m = 0.1$, and maximum generation be 150, the distribution diagram of the resulted solutions of 20 independent runs is illustrated in Fig.1. Secondly, use 10 independent simulations for performance estimation ($n = 10$), others are the same as stated above and the hypothesis test is still not applied, the distribution diagram of resulted solutions of 20 independent runs is illustrated in Fig.2.

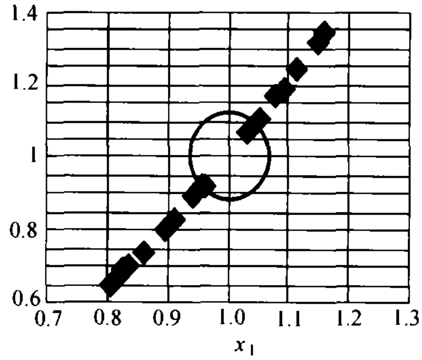

Fig. 1 Results of GA without hypothesis test ($n = 1$)

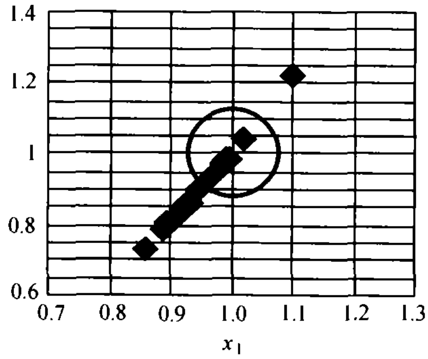

Fig. 2 Results of GA without hypothesis test ($n = 10$)

Then, we study the HTGA ($n = 10$) with other parameters as before, the distribution diagram of the resulted solutions of 20 independent runs is illustrated in Fig.3. If the noise magnitude increases to 0.05, the distribution diagrams of the resulted solutions of 20 independent runs with $n = 10$ and $n = 20$ are illustrated in Fig.4 and Fig.5 respectively. In addition, the performances of 20 independent runs of classic GA and HTGA under different noise magnitude are listed in Table 1, where $J_{\min}$ denotes the theoretically expected optimal value, $\bar{L}(x^*)$ and $\bar{J}(x^*)$ denote the average estimated performance and the average expected performance of the resulted solutions (evaluation without noise) respectively.

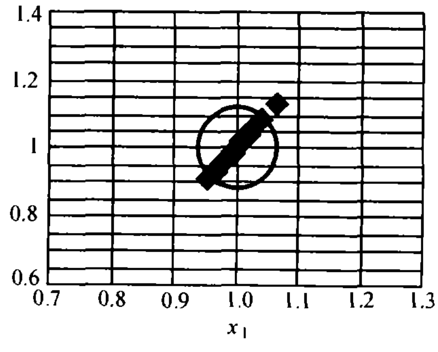

Fig. 3 Results of HTGA ($n = 10, \eta = 0.01$)

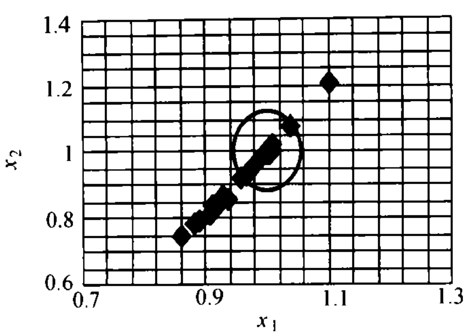

Fig. 4 Results of HTGA ($n = 10, \eta = 0.05$)

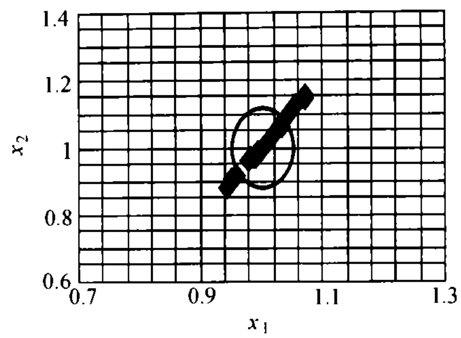

Fig. 5 Results of HTGA ($n = 20, \eta = 0.05$)

Table 1 Comparison between classic GA and HTGA for stochastic functional optimization

η	without HT $n = 1$		without HT $n = 10$		HTGA $n = 10$		HTGA $n = 20$		
	$\bar{L}(x^*)$	$\bar{J}(x^*)$	$\bar{L}(x^*)$	$\bar{J}(x^*)$	$L(x^*)$	$\bar{J}(x^*)$	$\bar{L}(x^*)$	$\bar{J}(x^*)$	J_{min}
$\eta = 0.01$	-0.0168	0.0160	-0.0046	0.0059	-0.0078	0.0013	-0.0057	0.0008	0
$\eta = 0.05$	-0.0381	0.0621	-0.0293	0.0244	-0.0390	0.0078	-0.0303	0.0033	

Firstly it can be concluded from the simulation results that more accurate estimation and better optimization quality can be achieved by increasing the evaluation number for stochastic optimization problem especially when noise magnitude is large (see the comparisons between Fig.1 and Fig.2, Fig.4 and Fig.5). Secondly it can be concluded that it can enhance the population diversity and reserve the best solution by incorporating the hypothesis test into the search procedure. This is helpful to avoid repeated search, premature convergence and being trapped in local minimum so that better and more robust performance can be achieved (see the comparison between Fig.2 and Fig.3, and Table 1).

4.2 Study on stochastic combinatorial optimization

Flow shop is a class of NP-hard combinatorial optimization problem with strong engineering background[6,7]. Due to the existence of non-deterministic factor, it is of more practicable significance to study stochastic scheduling problems, where flow shop problem with random processing time is a typical one[8]. Here, flow shop problem with random processing time uniformly distributed is considered. Let $p_{i,j}$ be the processing time of job i in machine j, which is subjected to uniform distribution $U((1-\eta)P_{i,j}, (1+\eta)P_{i,j})$, where $P_{i,j}$ is the expected processing time, and η is noise magnitude. We select three benchmarks from [9] for performance testing, namely Rec07 (20 jobs, 10 machines), Rec19 (30 jobs, 10 machines), and Rec25 (30 jobs, 15 machines) whose processing data are used as $P_{i,j}$. Let $\eta = 0.15$ and makespan [6] be the objective for the following simulation.

Because of the complexity of flow shop solving, local search ability of GA may be weakened to some extent by incorporating hypothesis test to enhance the population diversity. To. compromise the exploration and local exploitation ability, in HTGA such a process is repeated that the hypothesis-test based search is consecutively applied for M generations after classic GA is continuously applied for N generations. The statistical simulation results of HTGA and classic GA (say SGA below) each with 20 independent runs for every instance are shown in Table 2 (LOX crossover[7], SWAP mutation[7], $P_s = 60, p_m = 0.1$, initial population randomly generated, maximum generation is 850, $n = 20, M = 100, N = 30$). In Table 2, $J(x^*)$ denotes the best performance evaluated with expected processing time among 20 resulted solutions, and others are the same as before.

Table 2 Comparison between classic GA and HTGA for stochastic combinatorial optimization problem

problem	J_{min}	SGA			HTGA		
		$\bar{L}(x^*)$	$\bar{J}(x^*)$	$J(x^*)$	$\bar{L}(x^*)$	$\bar{J}(x^*)$	$J(x^*)$
Rec 07	1566	1577.7	1596.9	1584	1568.2	1584.3	1566
Rec 19	2093	2152.6	2171.2	2141	2135.7	2152.7	2115
Rec 25	2513	2611.0	2629.5	2588	2591.1	2608.9	2565

It is shown from Table 2 that the average estimated performance, average expected performance and best expected performance of HTGA are obviously better than those of classic GA without hypothesis test when solving flow shop with random processing time, even with the same evaluation replications. Moreover, the robustness of HTGA is demonstrated since the expected performance of the resulted solution is very close to the theoretically optimal value. So, once again it is shown that by suitably incorporating hypothesis test into GA, HTGA is powerfully able to solve complex stochastic combinatorial optimization problem in order to achieve good optimization quality and robust performance.

5 Conclusion

This paper proposed a hypothesis-test based GA to solve stochastic optimization problems with non-deterministic and multi-modal properties. By applying performance estimation and hypothesis test based on multiple simulations, it can enhance population diversity and decrease repeated search so as to be helpful for GA to effectively solve stochastic problems. The effectiveness and robustness of the proposed algorithm have been demonstrated by simulation results based on stochastic functional and combinatorial optimization. The future work is to theoretically study the convergence property, design adaptive mechanism for parameters, and apply for other complex scheduling problems.

References:

[1] WANG L, ZHANG L, ZHENG D Z. Advances in simulation optimization [J]. *Control and Decision*, 2003, 18(3): 257 - 262, 271. (王凌,张亮,郑大钟.仿真优化研究进展[J].控制与决策, 2003, 18(3): 257 - 262, 271.)

[2] WANG L. *Intelligent Optimization Algorithms with Applications* [M]. Beijing: Tsinghua University & Springer Press, 2001.

[3] MICHALEWICZ Z. *Genetic Algorithm + Data Structures = Evolution Programs* [M]. Berlin: Springer-Verlag, 1996.

[4] BEYER H G. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice [J]. *Computer Methods in Applied Mechanics and Engineering*, 2000, 186(2/4): 239 - 267.

[5] LIU X. *Introduction to Statistics* [M]. Beijing: Tsinghua University Press, 1998.

[6] PINEDO M. *Scheduling: Theory Algorithms and Systems* [M]. Engle-wood Cliffs, NJ: Prentice-Hall, 1995.

[7] WANG L, ZHENG D Z. An effective hybrid heuristic for flow shop scheduling [J]. *Int J of Advanced Manufacturing Technology*, 2003, 21(1): 38 - 44.

[8] BALASUBRAMANIAN J, GROSSMANN I E. A novel branch and bound algorithm for scheduling flowshop plants with uncertain processing times [J]. *Computers and Chemical Engineering*, 2002, 26(1): 41 - 57.

[9] REEVES C R. A genetic algorithm for flowshop sequencing [J]. *Computers and Operations Research*, 1995, 22(1): 5 - 13.

作者简介:

张 亮 (1982 —),男,清华大学自动化系研究生,主要研究智能和随机优化理论与方法,E-mail: zhangliang02@mails.tsinghua.edu.cn;

王 凌 (1972 —),男,1999 年在清华大学自动化系获博士学位,现为该系副教授,研究兴趣为优化理论与方法、生产系统建模与调度,已出版专著 2 部,发表论文 80 余篇,E-mail: wangling@tsinghua.edu.cn;

郑大钟 (1935 —),男,自动化系教授,博士生导师,主要研究兴趣为 HDS 与 DEDS 的建模、控制与优化.