The Empirical Study of the Schema Theory of Genetic Algorithm Based on 3-satisfiability Problem

ZHANG YuAn[1,a], LI BingFen[,2,b],

[1]Department of Computer Technology and Applications,Qinghai University, Xining 810016 China

[2]School of Mechanical Engineering,Qinghai University, 810016, China

[a]2011990029@qhu.edu.cn, [b]734034138@qq.com

Keywords: schema theory; genetic algorithm; 3-SAT problem

Abstract. Undoubtedly, schema theory is one of the most significant of genetic algorithm theory. Many research and application based on the schema theory of genetic algorithms have been reported. Through mathematical analysis of genetic algorithm, every genetic operator all has great influence on genetic algorithm and the schema of genetic algorithm. So, this article studied the experimental of genetic algorithm based on 3-SAT problem to analyze the influence of the initial population, genetic operators on the schema.

Introduction

NP-Completed problem is an important problem having many real world applications which is hard to deal for conventional computers[1]. Many efforts during recent years have been done to solve NP-Complete problems, but no efficient solution is found yet[2]. The satisfiability problem is a well-known NP-Complete computational problem or which the fastest known sequential algorithms require exponential time[1], as well as is the first proven for NP-Complete problems[3].

3-SAT problem has the huge search space, therefore, deterministic approaches are not recommended or optimizing of these unctions with a large number of variables. In contrast, an evolutionary approach such as PSO may be applied to solve these kinds of problem, effectively[4]. There are many genetic algorithms for solving 3-SAT problem. Schema theories are among the oldest, and probably the most well-known, classes of models of evolutionary algorithms[5]. Many scholars have studied the schema theory of the genetic algorithm about binary coding in detail, and make it more perfectly[6]. In this paper, we study schema of 3-SAT problem based on genetic algorithms, and study the influence of the initial population, genetic operators on the schema.

The 3-SAT Problem and Genetic Algorithm

In this section, description of the 3-SAT problem is presented. A Boolean satisiability problem (SAT) involves a Boolean formula F which consists of a set of n Boolean variables (their value either be True or False) $x_1, x_2, ..., x_n$. The formula $F = C_1 \wedge C_2 \wedge ... \wedge C_m$, in conjunctive normal form (CNF), is a conjunction of m clauses. The $\wedge$ represents "and" operation . If a SAT problem that its each clause C of length is k, then means it is a k-SAT problem. C_i of length k is a disjunction of k literals, $C_i = l_i1 \vee l_i2 \vee ... \vee l_ik$, where a literal is either a variable x_j or its negation $\neg x_j$ $(1 \leq j \leq n)$. The $\vee$ represents "or" operation. It has been proven that the k-SAT problem is NP-Complete or any $k \geq 3$[7]. Throughout this paper, n and m will represent the number of variable and that of the clauses of the input formula F. For constant k, in this paper, we just consider $k = 3$, meanwhile, we only concerned with the 3-SAT problem in which each clause consists of just 3 literals. A formula F is satisfiable if there is a truth assignment to its variables satisfying every clause of the formula, otherwise the formula is unsatisfiable[4]. The k-SAT problem is a problem to study whether there exists an assignment of truth values to a set of Boolean variables that satisfy a conjunctive normal form formula to be true[8].

In recent decades, there are many practical applications and research about the 3-SAT have been reported. Genetic algorithms (GAs) are random heuristic search (RHS) algorithms with a wide range of applications in adaptation and optimization problems. Genetic algorithms mimic the basic mechanisms of biological evolution and molecular genetics in simplified form[9]. Genetic algorithms

work over a set of potential solutions by simulating the process of biologic evolution to search for the best solutions[10]. It based on populations of individual solutions which evolve according to some stochastic operators like selection, mutation and crossover. Fitness of the individual to the environment is determined by the quality of the solution encoded by the individual, and it is the important indicator to determine whether the individual survival. Since genetic algorithms were proposed by Holland and his students[11] in 1975, they have been used in wide range of applications, e.g. for the search problem as well as numerical and combinational optimisation problems[12].

In the next, we give a brief introduction to genetic algorithms.First, some concept is introduced. In a 3-SAT problem based on genetic algorithms, the possible solution of 3-SAT problem is called chromosomes.The GA encoding includes binary, gray,and decimal encoding, etc. In this experiment, the binary encoding is used. A chromosomes (a sequence of Boolean) is composed of genes, in this paper a gene means a variable (a literal). A population is a set of all chromosomes, so a chromosome can be called a individual. The performance of a individual is the fitness of the chromosome. The population and the encoding formation are given by:
101000...100011 $chromosome_1$ (n represents the number of the sequence of chromosome)
.....
111000...111100 $chromosome_M$ (M represents the number of the whole population)

The fitness function expression is described as:

$$f(t) = \sum_{i=1}^{m} C_i, \tag{1}$$

Where C_i refers to the true or false of the i clause of SAT problem. m refers to the number of clauses of the SAT problem. When all the clauses all satisfied, the value of fitness is m, then the optimal solution is found, and it means the maximum fitness is m.

The ranking, selection and copy:

Selection will give the better individuals a bigger chance to survive. Ranking is means that chromosomes are sorted by the fitness values of each chromosome, from the best to worst. In this design, a half of better individuals are selected, copied and survived.

The crossover and mutation:

A conventional exchange way is used in this paper. Each of chromosomes has the same uniform probability to exchange and form the new individual.

Mutation operation simulates biological gene mutation which caused by change of natural environment. In this design, each gene of each chromosome is in the same probability of mutation. Generally, mutation rate is very small.

The Termination Condition:

If the chromosome which satisfies all the clauses have been found, then the algorithm goto the end. Otherwise, the algorithm keeps running until it reaches the preset fixed threshold number of algebra[13].

The Algorithm Implementation

a.Initialize, get initial population.

b.To evaluate the initial population by fitness function.

c.To determine whether the optimal solution is found or a termination condition is reached. If yes, break; else continue.

d.Crossover.

e.Mutation.

f.Selection.

g.To evaluate the new population by fitness function.

h.To determine whether the optimal solution is found or a termination condition is reached. If yes, break; else continue the step d.

Next, the principle of GA is applied in 3-SAT problem is introduced. Initially, a random population is created. The fitness of initial population is assessed according to the fitness function. The fitter individual is chosen as parents which is exactly like a natural creature which obeys the rule of Darwinian evolution theory. The new population get from the selected individuals who follow the genetic

process of evaluation, order, selection, copying, crossover and mutation. The new population is then evaluated again. This cycle continues till suitable solution found. The suitable solution is called the optimal solution[14].

Basic of Concept for Schema

In this section we provide some basic of concept of schema. The primary important notion is the schema. A schema is a similarity template that represents a set of solutions in genetic algorithms. For binary coding, a schema is usually a string of symbols form the alphabet { 0, 1, * },where the character * is interpreted as a "0 or 1 $is\ all\ ok$" symbol and it means "$don't\ care$" state. For example, "01*00*" bits string has 4 states including: 010000, 011000, 010001 and 011001[15-16].

Generally, Schema is a frame of subsets of a chromosome that fixed section will be alike[16]. Notably, chromosomes with the same schema would have no improvement on each other information, but if schema is applied on a set of chromosomes with high fitness, this fitness will be remained in future algebra[17].

Definition 3.1

The number of fixed positions in the template H is defined as the order of the schema which written $O(H)$.For example:

$O(001**01*0) = 6$

Definition 3.2

The distance between the first and last specific positions in the template H is defined as the defining length which written $\delta(H)$.For example:

$\delta(001**01*0) = 8$

$\delta(001**01**) = 6$

Definition 3.3

According to the schema theorem, under the operators of the genetic operators such as selection, mutation and crossover, the schema with a low order, short defining length and its average fitness higher than the population average fitness will go exponential in the offspring.

Experimental Results

Generally, the representation of the problem solution that it can be called the structure of chromosomal, encoding scheme, the genetic operation such as crossover scheme, mutation scheme and selection strategy, all highly affects the speed of genetic algorithms. So, the structure of chromosomal, encoding scheme, the genetic operation such as crossover scheme, mutation scheme and selection strategy, all highly affects the schema of genetic algorithms. And in this paper, we study the influence of the initial population, crossover scheme and mutation scheme on the schema.

A set of benchmark instance was used in this study, those datum are retrieved at network database[18]. Numerical experiments of genetic algorithm based on 3-SAT problem with 20 variables and 91 clauses are performed. In this paper, all the experiments were performed repeatedly, and results were averaged over 10000 runs[19].

Next, some notions used in the following figures are introduced: $S(1)$ represents the number of fixed positions in the template is 1, it also means $O(H)$ = 1. $S(2)$ represents $O(H)$ = 2, then $S(k)$ represents $O(H)$ = k;Pc represents crossover rate, Pm is mutation rate;c. N represents population size, T is the algebra of the evolution.L represents the number of variables.

Figure 1 shows the different schema with N = 100, Pm = 0.02, Pc = 1, L = 20. At the beginning, we observe that the different schema is almost in proportion to the algebra. With the increasing of algebra, the schema trends to coverage. Comparing results with different schema ($s(k)$, k = 1,2,3,4,5,10,15,20), it is clear that the $O(H)$ is larger, the percentage of this schema is smaller.

Figure 2 shows the different population size with schema $O(H)$ = 10, N = 10, Pm = 0.02, Pc = 1, L = 20. From this figure, we observe that the population size N will influence the schema , but it not has influence the trend that with the algebra grow,the schema trends to coverage. The schema is

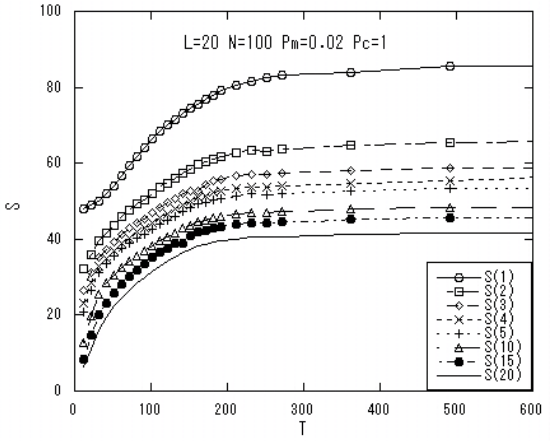

Fig. 1: Different schema with population size N = 100, Pm = 0.02, Pc = 1, L = 20. The horizontal axis represents T.

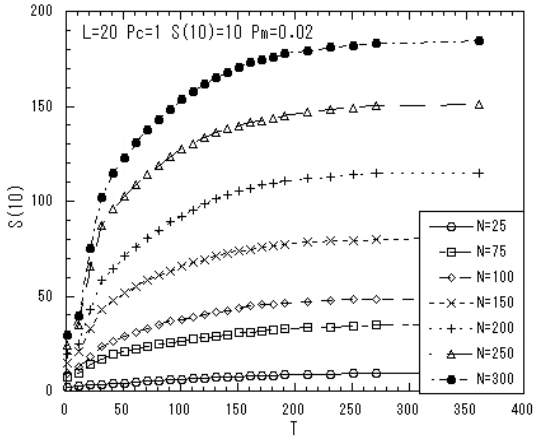

Fig. 2: The S(10) with respect to different population size N. The horizontal axis represents T.

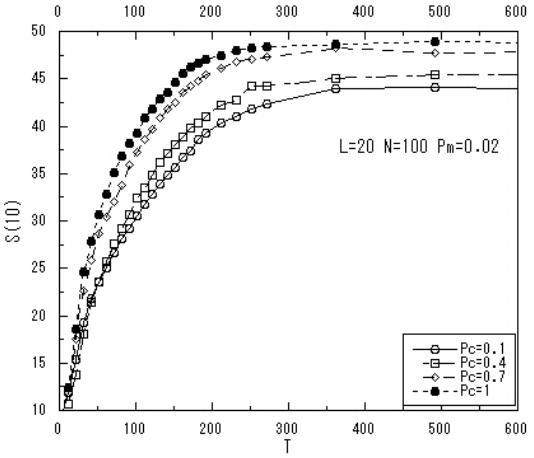

Fig. 3: The relationship between $S(10)$ and different Pc . The horizontal axis represents T.

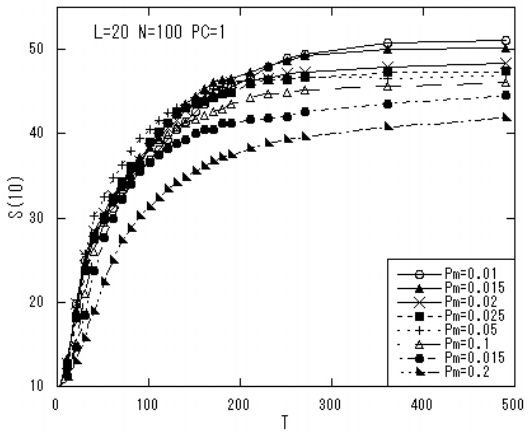

Fig. 4: The relationship between $S(10)$ and different Pm. The horizontal axis represents T.

almost in proportion to the algebra, and finally it trends to coverage. Comparing results with different population size, it is clear that the population size is larger, the number of the fixed schema is larger.

Figure 3 shows the different crossover rate with $O(H)$ = 10, N = 100, Pm = 0.02, L = 20. Comparing results with different crossover rate, it is clear that with the increasing of crossover rate, the percentage of this schema is lager. It means the crossover will benefit to survival of this schema.

Figure 4 shows the different mutation rate with $O(H)$ = 10, N = 100, Pm = 0.02, L = 20.Obviously, when the mutation rate more than 0.05 and less than 0.2, we can know that the mutation rate is larger, the percentage of this schema is smaller. When the mutation more than 0.05, the mutation rate inhibits the development of this schema. Figure 5 shows the curve of $O(H)$ = 10 with Pm = 0, 0.0005, 0.001, 0.005, 0.01. Clearly, when the mutation rate less than 0.01, then we can conclude that the mutation rate is larger, the percentage of this schema is larger. When the mutation rate less than 0.01, the mutation rate contributes to development of this schema.

Figure 6 shows the different curve of schema $O(H)$ = 10 at the algebra T = 1, 101, 191, 361, 491. Obviously, we observe that the schema is almost in proportion to the algebra. With the increasing of algebra, the percent of schema of $O(H)$ = 10 is increasing.Comparing the crave of different algebra, when Pm = 0.01, the percentage of this schema is largest.

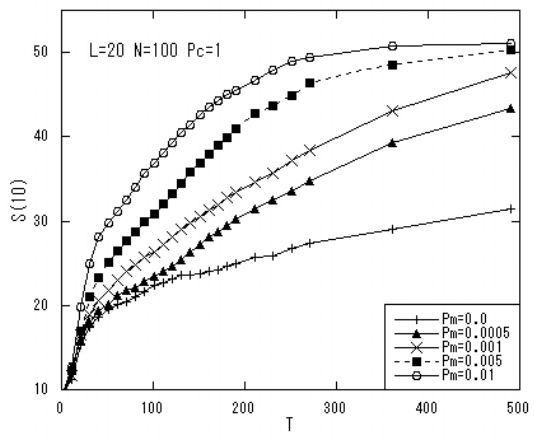

Fig. 5: Algebra-dependence of $S(10)$ with Pm = 0, 0.0005, 0.01, 0.005, 0.01.

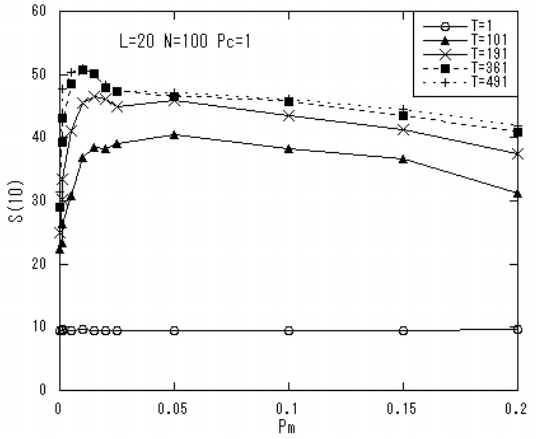

Fig. 6: Pm-dependence of $S(10)$ with T = 1, 101, 191, 361, 491.

Summary

The genetic algorithms parameters, such as, the number of variables, crossover rate, mutation rate and population size, all influence the schema of GA greatly. From those figure, we can conclude that:
1) We observe that the schema is almost in proportion to the algebra. With the increasing of algebra, the schema trends to coverage.Comparing results with different schema, it is clear that the $O(H)$ is larger, the percentage of this schema is smaller.
2) For population size, we observe that the population size N have an great influence the schema. With the growth of the population size, the degree of the increase of schema is larger.
3) For crossover rate, it is clear that with the increasing of crossover rate, the percentage of this schema is lager. It means the crossover will benefit to survival of this schema.
4) For Mutation rate, when the mutation rate less than 0.01, the mutation rate contributes to development of this schema.When the mutation rate more than 0.05 and less than 0.2, we can know that the mutation rate is larger, the percentage of this schema is smaller. When the mutation more than 0.05, the mutation rate inhibits the development of this schema. When Pm = 0.01, the percentage of this schema is largest.
5) From all the figures, we can concluded that the particular model will increase with the radio between it's average fitness value and the average fitness value of the population. In other words, those individuals that fitness values are higher than the average fitness value of the population will be more representative in the next algebra.

Acknowledgment

This work is supported by Natural Science Foundation of Qinghai Province(No.2013-Z-930Q),Natural Science Foundation of China(No.61363019).

References

[1] Braich Rs,Chelyapov N,Johnson C,Rothemund Pwk,Adleman L. *"Solution of a 20-variable 3-SAT problem on a DNA computer"*. Vol.296 no.5567(2002),p.499-502.

[2] Sama Goliaei,Saeed Jalil.*"An optical solution to the 3-SAT problem using wavelength based selectors"*.J Supercomput.Vol.62 no.2012 (2012),p.663-672.

[3] Cormen TH, Leiserson CE, Rivest RL, Stein C. *"Introduction to algorithms"*, 2nd edn. MIT Press,Cambridge(2001).

[4] Nasser Lotfi, Jamshid Tamouk, Mina Farmanbar. *"3-SAT Problem: A New Memetic-PSO Algorithm"*. Neural and Evolutionary Computing Eprint arXiv: 1306.5070.2013,06.

[5] R. Poli and N. F. McPhe. *"General Schema Theory for Genetic Programming with Subtree-Swapping Crossover: Part II"*. Evolutionary Computation. Vol.11, No.2(2003),p.169-206.

[6] Neubauer A. *"The circular schema theorem for genetic algorithms and two-point crossover "*.Genetic Algorithms in Engineering Systems Innovations & Applications .galesia.second in. (1997),p.209-214.

[7] Liu, Wenbin ; Liu, Xiangrong ; Wang, Shudong ; Xu, Jin ; Gao, Lin. *"Solving the 3-SAT Problem Based on DNA Computing"*. Journal of Chemical Information and Computer. Sciences,Vol.43, No2(2003), pp.1872-1875

[8] J. Gu. *" Parallel Algorithms for Satisfiability(SAT) Problem"*. DIMACS volume Series in Discrete Mathematics and Theoretical Computer Science, American Mathematical Society, Vol.22(1995),p.105-161.

[9] Andre Neubauer. *"Theory of the Simple Genetic Algorithm with α-Selection"*.Conference on Genetic & Evolutionary Computation, (2008),p.1009-1016.

[10] Enrigue Alba, Hugo Alfonso,Bernabe Dorronsoro.*"Advanced Models of Cellular Genetic Algorithms Evaluated on SAT"*. On Sat Acm Gecco, (2005),p.1123-1130.

[11] J. H. Holland. *"Adaptation in Natural and Artificial Systems -An Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence"*.First MIT Press Edition, Cambridge, 1992.

[12] M. Mitchell. *"An Introduction to Genetic Algorithms"*.MIT Press, Cambridge, 1996.

[13] Y.A Zhang,B.F Li,Q Meng,Q.Q Hu,Q.L Ma. *"The Experimental Analysis of the Computational Efficiency of Genetic Algorithms for 3-satisfiability Problem"*.The 11th International Conference on Natural Computation,icnc2015,vol.1(2015),p.247-250.

[14] Long wei-jun, Ben De, Bakhshi Asim D, Zhang gong. *"Pattern synthesis optimization of 3-D ODAR based on improved GA using LSFE method"*. Journal of Harbin Institute of Technology. Vol.1(2011),p.96-100.

[15] Chenxia Jin, Fachao Li, Marzana Wilamowska-Korsak, Ling Li and Liuliu Fu. *"BSP-GA: A new Genetic Algorithm for System Optimization and Excellent Schema Selection"*.Systems Research and Behavioral Science.Syst. Res.Vol.31(2014),p.337-352.

[16] Mahdi Jampour, Mahdi Yaghoobi and Maryam Ashourzadeh. *"Fractal Images Compressing by Estimating the Closest Neighborhood with Using of Schema Theory"*. Journal of Computer Science. Vol.6,No.5(2010),p.591-596.

[17] Mitra, S.K., C.A. Murthy and M.K. Kundu. *"Technique for fractal image compression using genetic algorithm"*. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society.Vol.7,No.4(1998),p.586-593.

[18] Information on http://www.cs.ubc.ca/ hoos/SATLIB/benchm.html

[19] Q.L.Ma, Y.A.Zhang, M.Sakamoto, H.Furutani. *"Study of computational performance of Genetic Algorithm for 3-satisfiability problem"*.The Seventeenth International Symposium on Artificial Life and Robotics 2012,Vol.1(2011),p.354-355.