Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

COPULA FOR ESTIMATION OF DISTRIBUTION ALGORITHM BASED ON GOODNESS-OF-FIT TEST

[1]JIA BAOJUN, [2]WANG LIFANG AND [3]CUI ZHIHUA

[1] Complex System and Computational Intelligence Laboratory, Taiyuan University of Science and Technology Taiyuan, 030024, China
[2] Complex System and Computational Intelligence Laboratory, Taiyuan University of Science and Technology Taiyuan, 030024, China
[3] Complex System and Computational Intelligence Laboratory, Taiyuan University of Science and Technology Taiyuan, 030024, China
E-mail: 344214701@qq.com , Wlf1001@163.com , cuizhihua@gmail.com

ABSTRACT

Inappropriate probabilistic model in the Estimation of Distribution Algorithm will reduce the quality of the solution, so goodness of fit test is introduced into Estimation of Distribution Algorithm to improve it. By means of test to select more suitable copula function under the statistical significance, that is to select more appropriate probability model under the statistical significance to solve the optimization problem. This modified Estimation of Distribution Algorithm improves the performance of the algorithm. The experimental results show the effectiveness of the proposed algorithm.

Keywords: *Estimation of Distribution Algorithm, Goodness of Fit, Probability Model, Copula.*

1. INTRODUCTION

Estimation of Distribution Algorithm (EDA) is an emerging theory in the field of evolutionary computation, which researches how to construct an appropriate multivariate joint distribution function. It is developed on the basis of the Genetic Algorithm. It creates a model of the probability distribution of the individuals in the solution space by means of statistical learning , then samples randomly according to the probability model to produce new population, and achieves finally the evolution of the population. With the rapid development of computer technology and information technology and with the more mature of the marginal distribution modeling problems, there will be more progress in this area in future [6, 13].

Estimation of Distribution Algorithm is mainly concentrated in the following areas: (1) Parallel Algorithm, that is the way of the use of parallelism in the modeling stage; (2) the combination of Estimation of Distribution Algorithm and other optimization algorithms; (3) the increase in the population diversity; (4) the theoretical research of Estimation of Distribution Algorithm, including its convergence analysis, time and space complexity analysis; (5) the application of Estimation of Distribution Algorithm.

Estimation of Distribution Algorithm is mainly divided into three categories: (1) Variables are mutually independent, UMDAc [12] and PBILc [14], which are represented , describe the probability model by using the normal distribution; (2) Variables are pair-wise correlations, MIMICc is one of the typical algorithms of Estimation of Distribution Algorithm, whose probabilistic model for each pair of variables is described by two-dimensional normal distribution; (3) Multiple variables are related, EGNA Algorithm is an Estimation of Distribution Algorithm based on a normal probability plot, MOA Algorithm [15] is an Estimation of Distribution Algorithm based on local Markov property , which samples under the undirected graph's local conditional probability. Zhong Weicai who proposed an Estimation of Distribution Algorithm [10] in 2005, its probability model is described by the general structure Gauss network, its advantage is without learning network structure and calculating the conditional probability density.

Wang Lifang and Zeng Jianchao proposed an Estimation of Distribution Algorithm based on the copula theory (cEDA) [6] in 2009, its research

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

object is binary copula function; its marginal distribution is Gaussian distribution. Zhang Jianhua, Zeng Jianchao proposed an Estimation of Distribution Algorithm Based on Sequential Importance Sampling [11] in 2010.

In fact, with the increase of the complexity of the problem involved, it is difficult to establish the precise probability model, but the inappropriate probability model will reduce the quality of the solution. It will inevitably become the bottleneck of the development of Estimation of Distribution Algorithm.

The paper is organized as follows. Section 2 provides a brief review on copula Estimation of Distribution Algorithm. Section 3 introduces the test algorithm. In section 3 we present a new algorithm. Section 4 conducts testing and analysis of the algorithm performance. Section 5 contains some concluding remarks.

2. COPULA ESTIMATION OF DISTRIBUTION ALGORITHM

In the copula theory, multivariate joint distribution can be broken down into two parts: (1) each variable marginal distribution function, (2) a copula function. Copula function is a multivariable function, in which each independent variable is in the range of [0, 1]. With the copula theory we can study marginal distribution and the correlation structure separately, and can choose the right distribution according to the actual situation. Copula function can completely describe the correlation between variables and can be used to construct more realistic multivariate probability distribution.

At the same time, in the course of the study we found the need for further research: in the Estimation of Distribution Algorithm of Archimedean copula, to what extent the parameter of the function affects the algorithm, as well as how to determine the appropriate parameter and appropriate copula function. The copula function affects the variable structure, and the parameter of the copula affects the variable structure, also affects the degree of fit between the actual probability distribution model and the estimated probability distribution model, and has an impact on algorithm performance further.

3. TEST ALGORITHM

To establish a probabilistic model of the solution space is the key in the Estimation of Distribution Algorithm. How to select a better model is very important to descripting the solution space appropriately. First, the optimization of Estimation of Distribution Algorithm is the high dimensional complex optimization problems of continuous domain. Accurate estimate of the probability distribution model, and macro control of the evolution of the population can improve the algorithm performance; Second, this is directly related to sampling according to the probability model selected, a suitable distribution should be easily sampled; Third, different algorithms are different in the optimization of the same standard test function. The difference has a certain relationship with goodness-of-fit between the probability model established and the actual probability model; Fourth, if the selected probability model can accurately describe the distribution of the solution space, the convergence of the algorithm can be certified.

Theoretically, for Estimation of Distribution Algorithm of infinite population size, if the probability model can accurately reflect the distribution of the selected population, then Estimation of Distribution Algorithm is global of convergence [9] under the three operators of the proportional selection, truncation selection, and two individual tournament selection.

This section proposes a copula Estimation of Distribution Algorithm based on the test of Sn (Sn-EDA). The probability model of the algorithm is to select the most suitable from the multiple alternative probability models. We create multiple alternative models first, then construct statistic Sn relying on the amount of structure of statistical test, finally select a more accurate reflection of the probability model of the selected population under the test of Sn. In the algorithm, we get next generation population according to the selected probability model above. We first get a population by sampling from the selected probability model, then we add some better individuals into the population according to the fitness values and add some mutation individuals into the population according to adaptive mutation operator, at last we consider the updated population as next generation population.

Journal of Theoretical and Applied Information Technology
31st January 2013. Vol. 47 No.3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Sn-EDA algorithm steps:

(1) Select k copula functions f1, …, fk as the research object, and these parameter values are t1, ..., tk;

(2) Initialization, that is to sample randomly in accordance with the uniform distribution to produce n individuals as the initial population and to determine the selectivity a and the mutation rate b;

(3) According to the fitness values, choose a * n individuals as a dominant population, and estimate the parameter values t1, ..., tk;

(4) In each generation evolution, sample from each copula function, then k alternative populations are obtained;

(5) Construct the statistic Sn, and select a relatively better structure from the alternative populations as a new population under goodness of fit test;

(6) Sample b* n individuals according with adaptive mutation operator , and add them into the new population as a mutation population;

(7) According to the fitness values, add M better individuals into the new population, considering this new population as a next generation population;

(8) If the algorithm termination condition is met, then the best fitness value of the individual is obtained; otherwise, turn to step 3 to continue.

4. TESTING AND ANALYSIS OF THE ALGORITHM PERFORMANCE

Under the overall framework of copula Estimation of Distribution Algorithm, clayton copula, gumbel copula, and frank copula are selected, and the marginal distribution function is normal distribution function. The optimization effect is researched based on the above algorithm.

Selection of the six test functions [5]:

Rosenbrock:

$$f(x)=\sum_{i=1}^{n-1}[100(x_{i+1}-x_i^2)^2+(x_i-1)^2]$$

$$-10<x_i<10$$

$$\min(f)=f(1,...,1)=0 \qquad (1)$$

Sumcan:

$$f(x)=-\{10^5+\sum_{i=1}^{n}|y_i|\}^{-1}$$

$$y_1=x_1,\ y_i=y_{i-1}+x_i \qquad i=2,...,d$$

$$-0.16\le x_i\le 0.16$$

$$\min(f)=f(0,...,0)=-10^5 \qquad (2)$$

Sphere:

$$f(x)=\sum_{i=1}^{n}x_i^2$$

$$-100\le x_i\le 100$$

$$\min(f)=f(0,...,0)=0 \qquad (3)$$

Schwefel 2.22:

$$f(x)=\sum_{i=1}^{n}|x_i|+\prod_{i=1}^{n}|x_i|$$

$$-10\le x_i\le 10$$

$$\min(f)=f(0,...,0)=0 \qquad (4)$$

Rastrigin:

$$f(x)=\sum_{i=1}^{n}[x_i^2-10\cos(2\pi x_i)+10]$$

$$-5.12\le x_i\le 5.12$$

$$\min(f)=f(0,...,0)=0 \qquad (5)$$

Griewank:

$$f(x)=1+\sum_{i=1}^{n}\frac{x_i^2}{4000}-\prod_{i=1}^{n}\cos(\frac{x_i}{\sqrt{i}})$$

$$-600\le x_i\le 600$$

$$\min(f)=f(0,...,0)=0 \qquad (6)$$

The parameter settings of 6 testing function are the same as the literature [6]: its population size is 2000, and its dimensions are 10, the test environment is in large samples and for small dimensions. In the choice of the dominant population, we select a part with truncation selection, we select another part with roulette-wheel selection, and the selection rate is 0.5. The maximum number of evaluation of the algorithm is 300,000 times, in addition we adopt the elitist strategy, every generation keeps 1% of the outstanding individuals in the preceding generation. Adaptive mutation operator is added

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

into the EDA, whose mutation rate is 0.05. Algorithm stopping criteria: (1) The results of the algorithm evolutionary with 25 consecutive generations are less than 1e-6; (2) Find the optimal value; (3) Maximum number of evaluation reaches to 300,000. To meet any one of three criteria, the algorithm is stopped. In this environment, each function independently runs 50 times and the results are as follows, in Table 1. The first two algorithm results are taken from the literature [6].

Table 1: Performance Comparison Of Clayton, Gumbel And Sn-EDA

Test Function	Algorithm	Mean	Mean Square Deviation
Rosenbrock	Clayton	8.3632e+000	6.07e-001
	Gumbel	6.6217e+000	8.46e-002
	Sn-EDA	6.5402e+000	7.29e-002
Sumcan	Clayton	-9.7367e+004	3.59e+003
	Gumbel	-9.0714e+004	4.25e+003
	Sn-EDA	-9.9828e+004	2.55e+002
Sphere	Clayton	1.3148e-007	1.45e-007
	Gumbel	3.5937e-009	1.94e-009
	Sn-EDA	1.2174e-009	1.78e-009
Schwefel 2.22	Clayton	3.0853e-005	1.22e-005
	Gumbel	1.4652e-007	7.19e-008
	Sn-EDA	4.9335e-008	1.97e-008
Rastrigin	Clayton	6.9989e-008	7.89e-008
	Gumbel	5.4907e-009	3.23e-009
	Sn-EDA	9.5183e-009	2.01e-008
Griewank	Clayton	3.6409e-007	3.29e-007
	Gumbel	9.4770e-009	6.43e-009
	Sn-EDA	5.5049e-010	3.88e-010

As can be seen from Table 1, except for the fifth test function, Sn-EDA algorithm is better than copula Estimation of Distribution Algorithm, has a better performance in the optimization results.

For the rosenbrock function, three algorithms converge to a local optimal solution, three algorithms achieve maximum evolution of the number of times; for the sumcan function, three algorithms achieve maximum evolution of the number of times, Sn-EDA is closer to the global optimal solution; for the remaining four test functions, the three algorithms converge to the global optimal solution, Sn-EDA has the highest accuracy.

5. CONCLUSIONS

In this paper we propose a copula Estimation of Distribution Algorithm based on the test of Sn. This Algorithm is introduced to exploit multivariate numeric optimization problem. In Estimation of Distribution Algorithm, to establish a probabilistic model of the solution space is the key. By means of test we select more appropriate probability model under the statistical significance to solve the optimization problem. We compare the new algorithm with the Archimedean copula EDAs. The results from Table 1 clearly demonstrate the advantages of the new Algorithm.

The results from Table 1 are obtained based on the Archimedean Copulae, so a great challenge would be to try to implement the tests based on copulae other than Archimedean Copulae; a greater challenge would be to try to implement the tests in small samples and for large dimensions.

ACKNOWLEDGMENTS

Many thanks to the Youth Research Fund of Shanxi Province (No. 2010021017-2), Supported by Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (No. 2010015) and the Doctor Fund of Taiyuan University of Science and Technology (No. 20122009) as well as the Excellent Graduate Innovative Projects of Shanxi Province (No. 20113121) for the financial support.

REFERENCES

[1] J. D. Fermanian, D. Radulović, M. H. Wegkamp. Weak convergence of empirical copula process. Bernoulli, Vol. 10, No. 5, 2004, pp. 847-860.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[2] J. D. Fermanian. Goodness-of-fit tests for copula. Multivariate Anal, Vol. 95, No. 11, 2005, pp. 119-152.

[3] C. Genest, B. Rémillard. Validity of the parametric bootstrap for goodness-of- fit testing in semiparametric models. Ann Inst H Poincaré Probab Statist, Vol. 44, No. 6, 2008, pp. 1097-1127.

[4] C. Genest, B. Rémillard, D. Beandoin. Goodness-of-fit tests for copulas: A review and a power study. Insurance: Mathematics and Economics, Vol. 44, No.2, 2009, pp. 199-213.

[5] W. Dong, X. Yao. Unified Eigen Analysis on Multivariate Gaussian Based Estimation of Distribution Algorithms. Information Science, Vol. 178, No. 15, 2008, pp. 3000~3023.

[6] L. F. Wang. Estimation of Distribution Algorithm based on the copula theory [M]. China Machine PRESS, 2012, in Chinese.

[7] S. D. Zhou, Z. X. Sun. A Survey on Estimation of Distribution Algorithms [J]. ACTA AUTOMATICA SINICA, Vol. 33, No. 2, 2007, pp. 113-124, in Chinese.

[8] P. Larrañaga, J. A. Lozano. Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation [M]. Kluwer Academic Publishers, 2002.

[9] Q. Zhang, H. Muhlenbein. On the convergence of a class of estimation of distribution algorithms IEEE Transactions on Evolutionary Computation, Vol. 8, No. 2, 2004, pp. 127-136.

[10] W. C. Zhong, J. Liu, F. Liu, et al. Estimation of Distribution Algorithm Based on Generic Gaussian Network [J]. Electronics and Information Technology, Vol. 27, No. 3, 2005, pp. 467-470, in Chinese.

[11] J. H. Zhang, J. C. Zeng. Estimation of Distribution Algorithm Based on Sequential Importance Sampling [J]. Computer Research and Development, Vol. 47, No. 11, 2010, pp. 1978-1985, in Chinese.

[12] P. Larrañaga, R. Etxeberria, J. A. Lozano, et al. Optimization in Continuous Domains by Learning and Simulation of Gaussian Networks[C]. In: Proceedings of the GECCO-2000 Workshop in Optimization by Building and Using Probabilistic Models. San Francisco, 2000, pp. 201-204.

[13] L. F. Wang, X. D. Guo, J. C. Zeng, Y. Hong. Copula Estimation of Distribution Algorithm Based on Exchangeable Archimedean Copula. Int. J. Computer Applications in Technology, Vol. 43, No. 1, 2012, pp. 13-20.

[14] M. Sebag, A. Ducoulombier. Extending population-based incremental learning to continuous search spaces. In: Back Th, Eiben G , Schoenauer M, Schwefel H P, editors, Proceedings of the 5th Conference on Parallel Problem Solving from Nature-PPSN V, Springer-Verlag, 1998, pp. 418-427.

[15] S. Shakya, R. Santana. A Markovianity Based Optimization Algorithm[R]. Spain: Department of Computer Science and Artificial Intelligence, University of the Basque Country, 2008.