Marginal Distribution in Copula Estimation of Distribution Algorithm Based Dynamic K-S test

Zhao Hui[1], Wang Lifang[2]

[1,2]Complex System and Computational Intelligence Laboratory, Taiyuan University of Science and Technology, Taiyuan, Shanxi, 030024, China

Abstract

Estimation of distribution algorithms based on copula, and a number of different distribution functions were selected as the dimension of the marginal distribution function by applying K-S test. Dynamic K-S test, that is, dynamic adjust frequency of K-S test in different stages. In the estimation of the probability model stage, according to the actual distribution of the dominant group respectively subject to inspection of the distribution function of each dimension. In the parameter estimation stage, according to fix the copula function parameters to make the emulation experiment. According to the obedience of the different marginal distribution function to sample separately, thus increasing the diversity of the population, and improving an execution efficiency of the estimation of distribution algorithms based on copula.

Keywords: *Estimation of Distribution Algorithms based on copula (copula EDA); Dynamic K-S test; the marginal distribution function*

1. Introduction

Estimation of Distribution Algorithms (EDA) [1] are a well established paradigm in Evolutionary Computation (EC),and are developed on the basis of Genetic Algorithm (GA).By the way of the traditional method, over and over again, this class of evolutionary algorithms builds probabilistic models for describing solutions' distribution from the community, and then to generate the new population by random sampling from the probabilistic models, realizing of population evolution until termination condition is satisfied.

Estimation of Distribution Algorithms based on copula (copula EDA) [2] are main to solute that how copula theory can be used in EDA and sample. In the light of copula theory, joint distribution is divided into margins and a copula. Obviously it is simpler to estimating the univariate margin than to estimating the joint, and the distribution model of dominant population are estimated more exactly, as well as, improving the convergence efficiency. According to fix the copula parameters to make the emulation experiment. Through the simulation experiment, this algorithm has better optimal value and can quickly converge to the optimal value, and finally the population has relative stability.

Appling K-S test to copula EDA, it give full consideration to the actual sample distribution, each dimension has its distribution function, it can be the same or different, so the use of K-S examination to determine the marginal distribution of dimensions can increase the diversity of population and improve execution efficiency.

Realization of K-S test [3] is that under the assumption in some generations (updated number of times:p1) , with the exception of each dimension of the marginal distribution in the next generation belong to the current generation of same distributions, only according to different parameters to determine the dimensions of different distributions. The follow ways are p1 is not a fixed number of times to realize K-S test, because in the early stage each dimensional distribution is very unstable, it can make the p1 values enough to small, that is to say, applying more K-S test to verify. On the contrary in the late stage, the dimensional distribution tends to be stable, the number of K-S test can be reduced accordingly, and the dynamic adjustment of K-S inspection frequency can give full consideration to the actual sample distribution, and has

relative flexibility.

2. Estimation of Distribution Algorithms based on Copula

2.1 A brief introduction of Estimation of Distribution Algorithms

EDA are the evolutionary model based on population, it mainly has the following three categories: variable independent; two variables are related each other; multiple variables are related. At present, some of the probability model can not make precise estimates between variables in the EDA, and when the problem size increases, the estimation of the joint distribution of the population will be very difficult [4]. The EDA of scientific value mainly embodied in three aspects: it is the biological evolution at the macro level of mathematical modeling, the probability model can describe the relationships between variables, effective solution to the nonlinear, variable coupling optimization problem, it is a new heuristic search strategy, and combined with statistical learning theory and stochastic optimization algorithm as well as the mixed design of other intelligent optimization algorithms, it greatly enriches the research content of hybrid optimization algorithm, and provides a new way of thinking for optimize algorithm.

2.2 A brief introduction of copula EDA

Copula theory is a new theory in the field of statistics; the research is how to construct appropriate joint distribution of multivariate, which has designated the univariate marginal distribution and correlations among multiple variables. EDA has the existence of the above problems, at the time of operation the copula theory is introduced to the EDA can simplify operate and decrease the time in estimating the probability distribution model.

Copula EDA is applied the copula theory to the EDA, which show both the characteristics of the copula theory and the advantages of the EDA, but also make up the problems in the EDA [5], copula theory will be isolated the marginal distribution from the complex joint distribution, and has the correlations between specified single variables marginal distribution function and multiple variables. In the light of copula theory, estimating joint distribution is divided into two parts, by the two step generate the joint distribution function, which is estimated becomes more simpler, and can better reflect the relationship among variables, meanwhile makes the complex problem simpler than before.

The earliest copula EDA is applied copula theory of two variables to the EDA[6]-[8], includes EDA based on copula thoughts begin to choose different of copula functions as the research object [6][7], then applied copula function to the MIMIC (Mutual information maximization for input clustering) algorithm [8] instead of conditional normal distribution, after, copula EDA select clayton copula functions and marginal distribution function of the empirical distribution [9][10] as references can be used in estimating the relationship among variables. As a result of normal distribution function of universality, and the function of normal distribution sampling is much simpler than the empirical distribution function sampling, then the function of normal distribution [2][11] applied to marginal distribution.

3. Copula EDA of Dynamic K-S test

3.1 K-S test selection

This section use the K-S test [12], reasons are as follow:
For the determination of the marginal distribution, in addition to mathematical model from practical experience, Pearson χ^2 test, the empirical distribution of type EDF test [13] and K-S test and other effective hypothesis testing method, are to test the validity of modeling assumption. But χ^2 test is used to discrete distribution, the study of continuity distributions need discrete, so some

useful information will be lost, directly lead to inspection effect is poor. The EDF test is only applied to small samples instead of complicating in the large sample. Therefore, we can use K-S test to test the validity of modeling assumption.

3.2 Simple introduction of K-S test

It is testing that a single sample is obey a certain predetermined hypothesis distribution or not, and test method is that compared with the cumulative frequency of sample data and specific theoretical distribution, if the gap is very small, the samples come from particular distributions.

Hypothesis testing problem is H0: samples from the overall distribution obey certain distribution; H1: samples from the overall distribution do not obey specific distribution.
F0(x) is to be assumed theoretical distribution, Fn(x) represents the random sample cumulative probability (frequency) function, hypothesis D=max|F0(x) - Fn(x) |.

Conclusion: when D > D (n, a), refused to H0, whereas accept H0 hypothesis. In which D (n, a) is a rejection threshold that significant level is a, and the sample capacity is n (it can be found in the tables).

3.2 Application of K-S test

The K-S test is applied to the copula EDA, used to test the samples' marginal distribution, which can choose variety of distribution function in statistics, such as the empirical distribution, normal distribution, exponential distribution, x^2 distribution. In this paper, clayton copula function as the link function and the number of K distributions are used the marginal distribution of the assumption. It is a kind of thought that according to the actual sample distribution and the K-S test to verify the specific distribution function, what's more determine the dimension of the marginal distribution function.

When the number of sample N<2000, W statistic of Shapiro-Wilk will test the suspect of hypothesis distribution; when the number of sample N>2000, D statistic of K-S can test hypothesis distribution. The number of sample is N=2000, and the application of K-S test, D statistic of K-S can test hypothesis distribution. When is tested, D for the test statistic get from calculation of samples, it compared the shape of samples distribution with hypothesis distribution to derive a value p (0<p<1, the actual significance level) to describe the degree of doubt about the idea. If the p< a, the original is assumed to be very suspicious, the data is not from the hypothesis distribution, conversely the data from hypothesis distribution.

3.2 Application of Dynamic K-S test in copula EDA

Add the K-S test to the copula EDA to determine edge distribution function as the preliminary work, Dynamic K-S inspection is dynamically adjusted K-S test frequency; p1 is not the fixed number of times. It can make K-S test algebraic obey a certain function, ($f(x)=k*x^q$, k=1,2...,q =1,2...), to determine times of the K-S test, and because of the unstable population distribution in the early period but relatively stable situation in the late, test frequency are becoming less and less during the whole period. x is running algebra of algorithm and f(x) is algebra of making K-S inspection. Thought is that: while is to achieve K-S test, can make each dimension of the marginal distribution of the next generation belong to the last generation of various distributions in some algebra (without K-S test algebraic), but according to different parameters to determine the dimension of different distribution .This can not only ensure the population stability, will not lose to diversity of population.

If k=1 and q=2, according to the function $f(x)=x^2$ to determine whether use the K-S test, if x=1,f(x)=1, test in the first generation, if x=2,f(x)=4, test in the fourth generation, meanwhile the second and third generation do

not test, and so on.

The basic algorithm described below:

Step1: Initialization population. Initialize randomly the population P0 with m individuals in the search space and set g←0.

Step 2: Calculate fitness value. According to the fitness function calculate the fitness value of the various points of the generation.

Step 3: Select a subpopulation. Select a subpopulation St of size s from P0 according to certain select-strategy (truncation selection, roulette selection).

Step 4: Select parameters of copula function. Select clayton copula functions as the link function, and select the parameter thita =1.

Step 5: Estimate probability distribution model of the dominant population. K-S test verify each dimension of the dominant population to obey probability distribution model, that is, each dimension of the dominant group as sample verify the specific probability distribution (F), it is mainly on K distribution function whether subject to normal distribution, Cauchy distribution, and t distribution for K-S testing. According to $f(x) = k*x^q$ determine whether the K-S test, x=1, 2..., then the generation of the corresponding K-S test is f(x). Set mark to flag=zeros (1, M), where M is the number of the dimension of distribution function.

Step 5.1: According to the dominant group get the sample mean and sample variance.

Step5.2: Compared with the selection of theoretical distribution function are obtained the corresponding value p (0<p<1).

Step 5.3: Selected significance level a, if all P values are greater than a, the samples do not obey the distribution selection, and not modify the tag, then structure the empirical distribution. Conversely it thinks that the sample obeys a certain distribution, and select the minimum distribution function as a marginal distribution, the dimension of the flag value are corresponding modified.

Step 6: According to the estimation of the probability model to sample [14], produce some new individuals. If the judgment of a dimension is not amenable to the selection of all the distribution, and structure the empirical distribution function to sample, otherwise make the corresponding distribution function of sampling. In the search space, according to the probability distribution F randomly generated l points as some individuals of the new generation.

Step 6.1: When make K-S test, according to the distribution of inspection to make sampling.

Step 6.2: When not make K-S test, in the generation check the marker flag of all dimensions, according to the different flag respectively corresponding to sampling.

Step 7: Update group. The s adaptive values in the g generation generate the better individuals, and the new individuals L, as well as PS-s-l individuals through mutation to form a new generation of groups, and set g ← g+1.

Step 8: If stopping criterion is not reached go to Step 2, the best individual in the g generation is the optimization results.

4. The performance test and analysis of algorithms

4.1Testing functions

In order to test the performance of the algorithm, and comparing the algorithm in the third section with copula EDA [15], using six test functions to test, where Rosenbrock and SumCan are prone to inappropriate covariance matrix, Sphere and Schwefel 2.22 are the two single peak function, Rastrigin and Griewank are two multi peak function.

Rosenbrock：

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

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

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

SumCan： $f(x) = -\{10^5 + \sum_{i=1}^{n} | y_i |\}^{-1}$ ， $y_1 = x_1$ ，

$y_i = y_{i-1} + x_i$， i=2，…，d， $-0.16 \le x_i \le 0.16$，
$\min(f) = f(0,...,0) = -10^5$.

Sphere： $f(x) = \sum_{i=1}^{n} x_i^2$，
$-100 \le x_i \le 100$，
$\min(f) = f(0,...,0) = 0$.

Schwefel 2.22： $f(x) = \sum_{i=1}^{n} |x_i| + \prod_{i=1}^{n} |x_i|$，
$-10 \le x_i \le 10$，
$\min(f) = f(0,...,0) = 0$.

Rastrigin：
$f(x) = \sum_{i=1}^{n} [x_i^2 - 10\cos(2\pi x_i) + 10]$，
$-5.12 \le x_i \le 5.12$，
$\min(f) = f(0,...,0) = 0$.

Griewank：
$f(x) = 1 + \sum_{i=1}^{n} \frac{x_i^2}{4000} - \prod_{i=1}^{n} \cos(\frac{x_i}{\sqrt{i}})$，
$-600 \le x_i \le 600$ ，
$\min(f) = f(0,...,0) = 0$.

Table 1: copula EDA and copula EDA of Dynamic K-S test in the fixed parameters comparison

Test function	Algorithm	Mean value	Variance
Rosenbrock	Copula EDA	6.5233	0.1359
	Copula EDA of Dynamic K-S test	7.0493	0.1992
SumCan	Copula EDA	-8.1066e+04	1.04e+04
	Copula EDA of Dynamic K-S test	-8.9611e+04	8.19e+03
Sphere	Copula EDA	4.6174e-08	2.61e-08
	Copula EDA of Dynamic K-S test	1.1553e-08	1.23e-08
Schwefel 2.22	Copula EDA	2.1417e-06	8.15e-07
	Copula EDA of Dynamic K-S test	5.4906e-07	6.00e-07
Rastrigin	Copula EDA	6.4494e-08	5.67e-08
	Copula EDA of Dynamic K-S test	2.5984e-08	4.95e-08
Griewank	Copula EDA	2.5861e-04	1.80e-03
	Copula EDA of Dynamic K-S test	7.8823e-08	7.81e-08

4.2Parameter setting

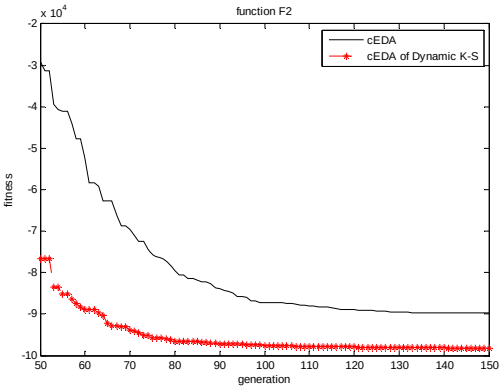
In the section, six functions are to be tested, the population size are 2000, the dimension are 10.when do select the dominant groups, the truncation as well as roulette wheel selection are part of the individual selection, and the rate of selection is 0.5. The number of times of appraise about maximum fitness value are 300000 in the algorithm, in addition to using the elite reservation strategy, every generation retain a generation in which at least 1% of the outstanding individual. In the algorithm it is used to mutation operator to choose outstanding individual of r

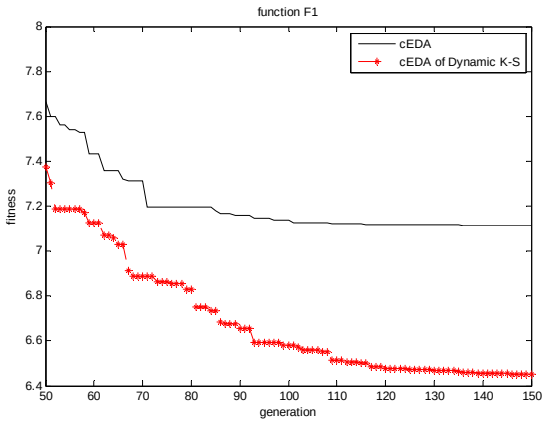
from the current population, then make incorporation variability, and add these individuals into the next generation, and the variation rate is 0.05.In addition, the significance level of K-S test is a= 0.05. The algorithm stopping criterion is: algorithm is less than 1e-6 in the 25 consecutive generations; find the optimal values; meet the maximum fitness value appraise number are 300000. Any one of three standards to meet, algorithm can stop. Each function is running independently the number of 50 in the different testing conditions. When make sampling, selecting fixed Clayton copula parameter thita is 1.

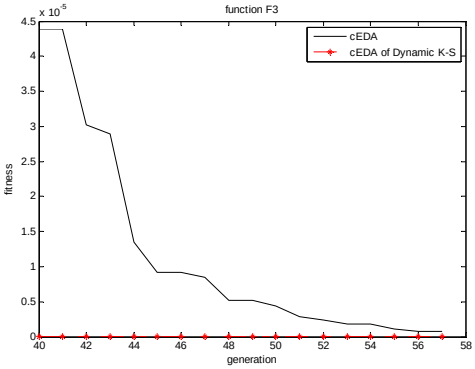
4.3 Results analysis

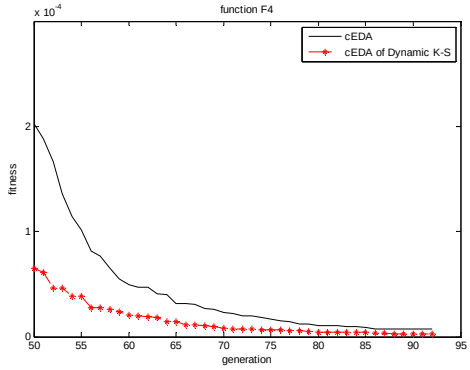
Experimental results can be seen from table 4.1, for the six test functions, when fix the parameter of clayton copula is 1, copula EDA of Dynamic K-S test is better than copula EDA in the optimization results, and fixing the parameter of the clayton copula make sampling is to save to estimate the parameters of time, improving the execution efficiency greatly.

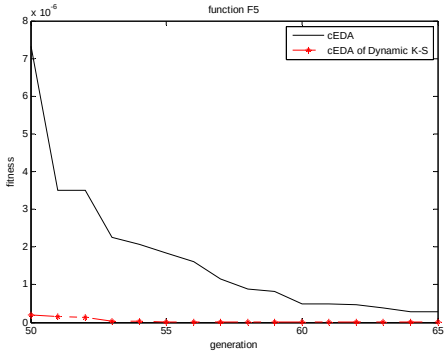
From the experimental results, we can see that, apply the K-S test to copula EDA is valid, the Dynamic K-S test are also played a better optimization effect, and compared to copula EDA optimization effect is obvious.











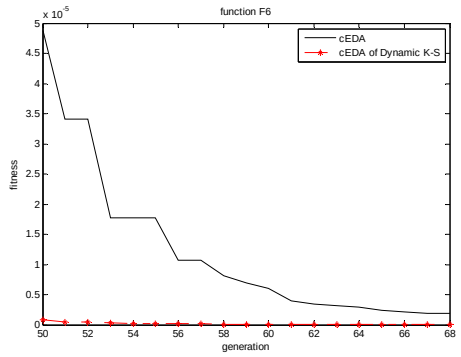


Figure 1 a running fitness of six testing functions by using two kinds of algorithm

Figure 4-1 shows the evolution diagram of six test functions in a running process, because the span of evolution chart data is bigger from the first generation to generation 150, which makes late graphic trend is not obvious, so we cut out the latter dozens of generations of evolution diagram. These diagrams illustrate the search process and the trend to search the optimal solution in later period by the copula EDA of Dynamic K-S test and copula EDA. It can be obviously observed the copula EDA of Dynamic K-S test can more quickly find the situation of optimal value than the copula EDA.

5. Conclusions

From the experimental results, we can see that, apply K-S test to the copula EDA is valid, and the Dynamic K-S test also played a better optimization effect, compared to copula EDA optimization effect is obvious.

1) Parameter choice. Research the performance influence of every parameters on the copula EDA of Dynamic K-S test, which parameters plays a key role, so as to guide how to choose appropriate parameters.

2) The analysis of algorithm convergence. According to the research between the copula theory and the existing distribution estimation algorithm convergence theory. Analysis the convergence of the copula EDA of Dynamic K-S test in theory.

3) Improve the algorithm in the late by local search method. Experiments show that the global detection ability of the copula EDA of Dynamic K-S test is strong, but local mining capacity is weak. Therefore, it can be cite other evolutionary algorithm thought in the copula EDA of Dynamic K-S test in order to enhance its local production ability.

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] Zhou S, Sun Z. The Overview of the Estimation of Distribution Algorithm, The Journal of Automation, 2007.2,33(2) : 113-124

[2] Wang L. The research of Estimation of Distribution Algorithm based on Copula, Doctoral Dissertation of Lanzhou University of Technology, 2011.4

[3] Hou S, Li Y, Liu G. K-S test application in fault diagnosis of bearings, coal mine machinery, 2004.10.11，（1）:129-130

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

[5] Wang L， Zeng J， Hong Y. Estimation of Distribution Algorithm Based on Copula Theory.In: IEEE Congress on Evolutionary Computation 2009. Trondheim, Norway, May 18-21, 2009. 1057-1063 (EI: 20094812504661)

[6] Wang L, Zeng J. Estimation of Distribution Based on Copula Theory. In: Exploitation of Linkage Learning in Evolutionary Algorithms，Ying-ping Chen (Ed.), Springer, 2010, 139-162

[7] Wang L，Zeng J，Hong Y. Estimation of Distribution Based on Archimedean Copulas. In: 2009 World Summit on Genetic and Evolutionary Computation. Shanghai, China, June 12-14, 2009. 993-996 (EI: 20093012219612)

[8]R.Salinas-Gutierrez,A. Hernandez-Aguirre,E.R.Villa-Diharce (2009).Using Copulas in Estimation of Distribution Algorithms In:LNAI 5845,MICAI 2009, A.Hernandez Aguirre et al.(eds.),pp:658-668

[9] Wang L, Wang Y, Zeng J, Hong Y. An Estimation of Distribution Algorithm Based on Clayton Copula and Empirical Margin. In: Life System Modeling and Intelligent Computing,Communications in Computer and Information Science, Wuxi，China，September 17-20， 2010, 98(1):82-88, DOI: 10.1007/978-3-642-15859-9_12. (EI: 20104513368882)

[10] Wang L，Zeng J，Hong Y. Using Gumbel Copula and Empirical Marginal Distribution in Estimation of Distribution Algorithm. In: 2010 Third International Workshop on Advanced Computational Intelligence (IWACI2010), Suzhou, China, August 25-27, 2010. 583-587. (EI: 20104613386525)

[11] Wang L，Zeng J，Hong Y. Using the Copula function to estimate the probability model and sampling in Estimation of Distribution Algorithm, Control and Decision, 2011.9，26（9）：1333-1338

[12] Wu Xizhi. Nonparametric statistics [M]. Beijing: China Statistics Press, 1992

[13] Stephens M A. Tests based on EDF Statistics. In Goodness-of- Fit Techniques, D, Agostino R B, Stephens, M. A. , edst Macel Dekker, New York, 1986 : 97-195

[14] Wang L，Zeng J，Hong Y，Guo X . Copula Estimation of Distribution Algorithm Sampling from Clayton Copula. Journal of Computational Information Systems, 2010, 6(7): 2431-2440

[15] Wang L， Guo X, Zeng J， Hong Y. Copula Estimation of Distribution Algorithm Based on Exchangeable Archimedean Copula. Int. J. Computer Applications in Technology(ISSN (Online): 1741-5047 - ISSN (Print): 0952-8091), 2012,43(1):13~20

Hui Zhao is a research student of Computer science and technology from Taiyuan University of Science and Technology. She received her BS degree in Computer science and technology from Hebei University of Engineering in 2009. Currently, her interests are in Intelligent Computing.

Liang Wang is an associate Professor in School of Computer Science and Technology at Taiyuan University of Science and Technology. She received her BS degree in Mathematics from Shanxi Normal University in 1998, received her MS degree in Computational Application from Taiyuan University of Science and Technology in 2005, and received her Dr. Degree in control theory and control engineering from Lanzhou University of Technology in 2011. Her research interests include intelligence computation and intelligence control.