



Article

Hybrid Approach with Improved Genetic Algorithm and Simulated Annealing for Thesis Sampling

Shardrom Johnson ^{1,2,3,*}, Jinwu Han ², Yuanchen Liu ⁴, Li Chen ¹ and Xinlin Wu ⁵¹ XianDa College of Economics and Humanities, Shanghai International Studies University, East Tiyyuhui Road 390, Shanghai 200083, China; 0710030@xdsisu.edu.cn² School of Computer Engineering and Science, Shanghai University, Shangda Road 99, Shanghai 200444, China; dewinhan@i.shu.edu.cn³ Information Centre, Shanghai Municipal Education Commission, Dagou Road 100, Shanghai 200003, China⁴ Faculty of Foreign Languages, Ningbo University, Fenghua Road 818, Ningbo 315211, China; 14111130@xdsisu.edu.cn⁵ Department of Education Evaluation Research, Shanghai Education Evaluation Institute, South Shaanxi Road 202, Shanghai 200031, China; xlwu@shhec.edu.cn

* Correspondence: jshardrom@shmec.gov.cn; Tel.: +86-021-5127-8051

Received: 11 July 2018; Accepted: 27 July 2018; Published: 30 July 2018

Abstract: Sampling inspection uses the sample characteristics to estimate that of the population, and it is an important method to describe the population, which has the features of low cost, strong applicability and high scientificity. This paper aims at the sampling inspection of the master's degree thesis to ensure their quality, which is commonly estimated by random sampling. Since there are disadvantages in random sampling, a hybrid algorithm combined with an improved genetic algorithm and a simulated annealing algorithm is proposed in this paper. Furthermore, a novel mutation strategy is introduced according to the specialty of Shanghai's thesis sampling to improve the efficiency of sampling inspection; the acceleration of convergence of the algorithm can also take advantage of this. The new algorithm features the traditional genetic algorithm, and it can obtain the global optimum in the optimization process and provide the fairest sampling plan under the constraint of multiple sampling indexes. The experimental results on the master's thesis dataset of Shanghai show that the proposed algorithm well meets the requirements of the sampling inspection in Shanghai with a lower time-complexity.

Keywords: sampling; genetic algorithm; simulated annealing algorithm; thesis and dissertation sampling; mutation strategy

1. Introduction

The thesis is one of the essential prerequisites for the graduation of postgraduates. Accordingly, thesis sampling inspection has been regarded as a significant task for analyzing and controlling the quality. Such a quality inspection ensures the high-quality standards of postgraduate education. Most of the existing research still is focused on the double-blinded review process of the degree theses. Moreover, there exists no applied research of the problem that Ph.D. dissertations and master's degree theses of the previous academic year must be sampled for inspection after [1] was issued.

In this paper, theses of the postgraduates who have achieved their degrees are used to carry out sampling inspection under the requirements of the Shanghai Academic Degrees Committee. The Shanghai Academic Degrees Committee builds the platform for provincial degree and postgraduate education, and the system is practically used in the Shanghai Postgraduate Education Quality Informatization Evaluation System [2].

To this end, this paper focuses on the sampling method that meets the actual need for quality inspection. The hybrid approach with the combination of the genetic algorithm and the simulated

annealing algorithm automatically samples master's theses from Shanghai in the previous year under the requirements of national sampling inspection of the Shanghai Academic Degree Committee. Development tools are used to implement the method to rebound the theoretical research and practical applications from the aspect of specific operations.

With the combination of the traditional genetic algorithm and the simulated annealing algorithm, a new mutation operation is also developed, leading to efficient and effective optimization for the sampling inspection of master's degree theses. Under the process of the method, the sampling inspection of the master's theses of provincial degree committees will be more standardized and institutionalized.

The rest of the paper is organized as follows: Section 2 introduces the background of paper quality sampling, sampling rules and related algorithms. The details of the hybrid algorithms with the improved genetic algorithm and the simulated annealing algorithm are given in Section 3. Experimental settings and dataset preprocessing are provided in Section 4. Section 5 presents the experimental results with the analysis. Section 6 draws the conclusion.

2. Background

2.1. Thesis Inspection via Sampling

Nearly 40,000 postgraduates graduate from Shanghai's universities each year. Because of the importance of the thesis to the graduation mentioned above, the quality of their theses needs to be controlled. For this reason, the Academic Degrees Committee of the State Council and the Ministry of Education formally promulgated [1] on 29 January 2014, which stipulates that the state will sample the doctoral dissertations and master's degree theses in the previous academic year nationwide. The sampling results would be used as an important indicator for the evaluation of degree authorization points. Rather than evaluating all items, sampling inspection aims to randomly select a small number of samples from a batch of individuals or a process according to an established sampling plan and judge whether the batch of individuals or the process meets the quality requirements [3]. It is an important method to describe the population, which has the features of low cost, strong applicability and high scientificity.

Compared to the national sampling after degree-granting, the Shanghai Academic Degree Commission has spent 300,000 to 400,000 RMB since 1997 for the random sampling and "double-blind" review of doctoral dissertations and master's theses before degree granting. It has directly promoted the overall attention of graduate students and their supervisors to the quality of their theses or dissertations [4]. Jiangsu, Hunan, Shandong and other provinces and cities have also followed suit and thus have some achievements. To a certain extent, they have all together exerted a positive effect on rectifying academic plagiarism and other unhealthy academic tendencies, just as the saying goes, "If you want people to think correctly, the first thing is to eliminate academic misconducts. When faults occur, then right or wrong is reversed" [5].

The primary purpose of this national sampling after degree-granting is not to influence the normal learning and employment of the applicants for degrees and the degree-granting of degree-cultivation units, but to promote the construction of professional disciplines and the training of talents. Emphasis is laid upon the performance of supervisors and the training quality of degree authorization institutions [6]. Moreover, [1] further clarified as the method for the sampling of master's dissertations is determined by provincial degree committees and the People's Liberation Army Academic Degree Committee.

Therefore, the correct understanding of the spirit of the national sampling method and how to implement it in the administrative area of Shanghai have become some of the critical tasks at the moment.

2.2. Sampling Rules

In the Shanghai sampling method, the sampling rules are listed as below.

- (1) The sampling rate of masters' theses from each master degree conferring institution is about 5%.
- (2) About 10% of foreign masters' theses need to be extracted for sampling inspection.
- (3) The sampling rate of theses of masters whose tutor guided ten or more postgraduates in the same year is about 10%.

2.3. Genetic Algorithm

The genetic algorithm (GA) [7] is an evolutionary algorithm inspired by the principle process of natural selection and survival of the fittest. GA is commonly used to generate high-quality solutions to optimization and search problems by relying on bio-inspired operations such as mutation, crossover and selection [8]. As shown in Figure 1.

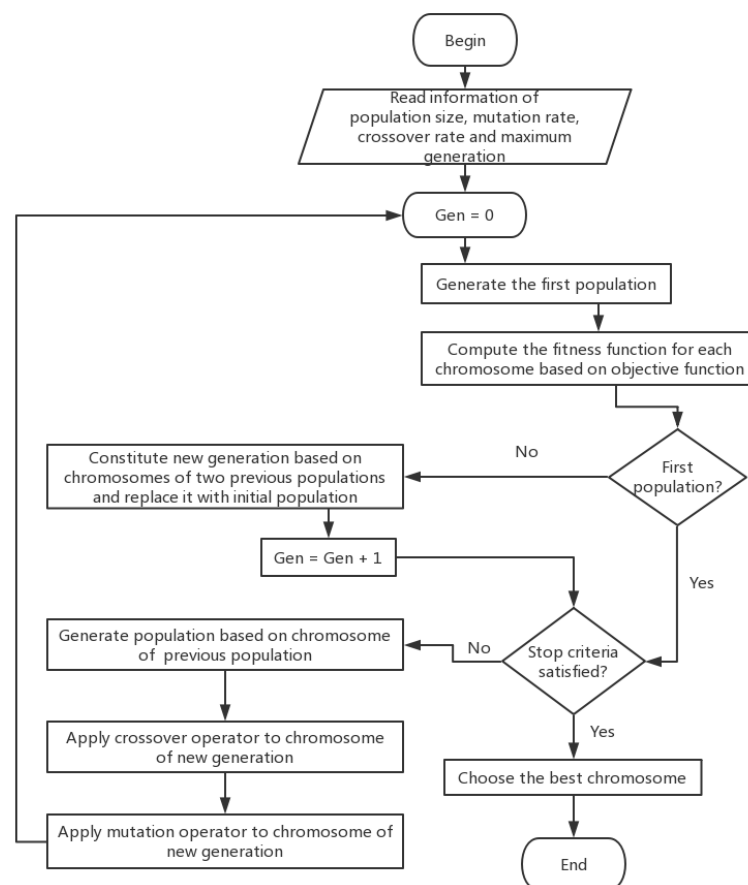


Figure 1. Schematic diagram of the genetic algorithm.

GA mainly borrows the principle of the survival of the fittest in biological evolution. First, the solution to the optimization problem is encoded. Here, the encoded solution is called a chromosome, and the elements that make up the encoding are called genes. Then, according to the objective function of the optimization problem, the corresponding fitness function is constructed, and the chromosome is selected by the probability distribution determined by the size of the fitness function. The survived chromosomes constitute a population, and at the same time, the parents are randomly generated according to the probability distribution. Then, the parents create offspring through mating between the codes, and consequently, the offspring mutate with a certain probability to form a new population. At this time, the decoding turns back to the solution to the optimization

problem. Finally, the termination rule is applied. Through the above cycle, the solution of the optimization problem is obtained.

Common iterative methods tend to fall into a local minimum trap, and thus, a dead loop phenomenon emerges, which makes the iteration process unable to continue. To overcome this shortcoming, GA adopts a global optimization approach [8]. Compared with traditional optimization methods (like enumeration, heuristics, etc.), GA has good convergence due to the adaptation of the biological evolution model [9]. GA also has several advantages as follows.

- It has no bias for specific problem areas.
- It has the capacity to perform a fast and random search.
- It uses a simple search process with the evaluation function.
- It features randomness as the probabilistic mechanism used for the iteration.
- It is extensible and easy to integrate with other algorithms.

Many variants of GA and its applications were thoroughly studied by [10]. Recently, GA has been one of the widely-used algorithms in solving optimization problems such as for image processing and communications. In this paper, we use it to make a sampling inspection of 40,000 masters' theses according to the three rules for thesis sampling in Shanghai mentioned in Section 2.2.

2.4. Simulated Annealing

Simulated annealing (SA) is a learning method that simulates the physical quenching process in thermodynamics. Its starting point is based on the similarity between the annealing process of solid materials in physics and general combinatorial optimization [11]. It is a greedy algorithm, but its search process introduces random factors. When iterating to update the feasible solution, a solution worse than the current solution is accepted with a certain probability, so it is possible to get the optimal global solution by jumping out of the local optimal solution [12]. In [13–15], the simulated annealing algorithm was improved or combined with other algorithms to achieve a better performance in experiments.

Figure 2 is an example. Assuming that the initial solution is the left blue point A, the simulated annealing algorithm will quickly search for the optimal local solution B, but after searching for the optimal local solution, it does not end there, but will move to the left with a certain probability. Maybe after a few of these non-locally optimal moves, the global optimum D is reached, and the local minimum jumps out.

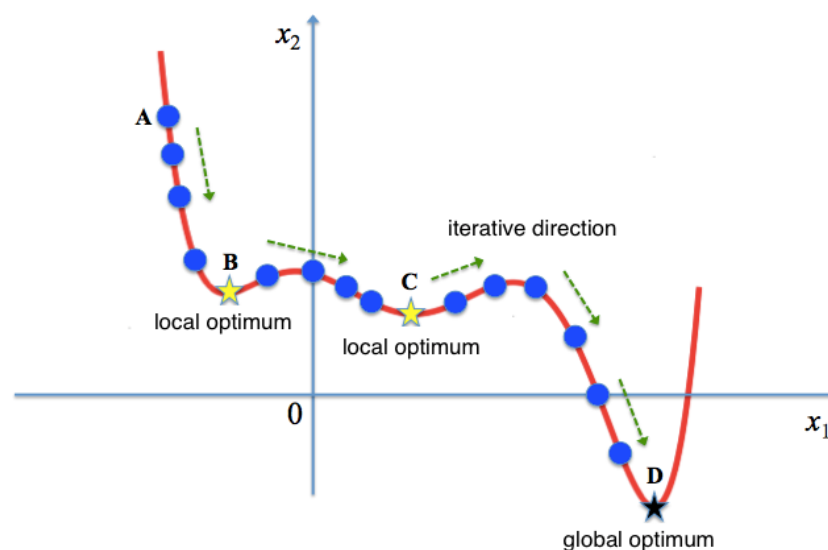


Figure 2. Simulated annealing algorithm.

According to the principle of thermodynamics, when the temperature is T , the probability of the temperature drop with the energy difference of dE is $p(dE)$, which is expressed as:

$$p(dE) = \exp\left(\frac{dE}{kT}\right), \quad (1)$$

where k is the Boltzmann constant; the value is $1.3806488 \times 10^{-23}$, and $dE < 0$. Therefore $dE/kT < 0$, so the value range of the $p(dE)$ function is $(0,1)$. In fact, the more intuitive meaning of this formula is: the higher the temperature, the greater the probability that the energy difference is $p(dE)$, and the lower the temperature, the lower the probability of temperature drop.

In practical problems, the calculation of “certain probability” here refers to the annealing process of metal smelting. Assuming that the currently feasible solution is x and the iteratively-updated solution is x_{new} , then the corresponding “energy difference” is defined as:

$$\Delta f = f(x_{new}) - f(x), \quad (2)$$

and its corresponding “certain probability” is:

$$p(\Delta f) = \begin{cases} 1, & \text{if } \Delta f > 0 \\ \exp\left(\frac{\Delta f}{kT}\right), & \text{if } \Delta f \leq 0 \end{cases} \quad (3)$$

Here, let $k = 1$. The simulated annealing algorithm can be described as below,

- Step 1: Initialization: initial temperature T (sufficiently large), original best solution x obtained after GA applied; solution obtained from the previous iteration x_{new} .
- Step 2: Calculate the increment $\Delta f = f(x_{new}) - f(x)$, where $f(x)$ is the optimization target. Here, it is the function that minimizes the mean absolute error ϵ calculated according to the sampling rules after the sampling result is obtained by decoding the solution. We use the result of one minus ϵ to represent the fitness function value; see Section 3.4.
- Step 3: If $\Delta f > 0$, x_{new} is accepted as the new current solution, otherwise x_{new} is accepted as the new current solution with the probability $\exp\left(\frac{\Delta f}{kT}\right)$.
- Step 4: If the termination condition is satisfied, the current solution is output as the optimal solution to terminate the program. Otherwise, wait for a new x_{new} to go to Step 2.

3. The Proposed Algorithm

3.1. The Main Idea

The major steps of the proposed algorithm are:

- Step 1: Initialize the variables of GA and SA; determine the initial temperature of annealing T_0 , temperature reduction parameter k , size of population s , crossover probability p_{cross} and mutation probability $p_{mutation}$ by the result obtained from the experiment in Section 4.2;
- Step 2: Randomly generate initial population p_0 , and encode each chromosome; the size of p_0 is 10;
- Step 3: Calculate the mean absolute error ϵ of the sampling scheme corresponding to each chromosome according to the sampling rules, then the fitness of each chromosome is denoted as f ; let $f = 1 - \epsilon$. Determine the fitness function of each chromosome in the population ($f = 1 - \epsilon$); see Section 3.4;
- Step 4: Roulette wheel method is used to select s chromosomes from p_0 to constitute new population p , the probability of choosing an individual depends directly on its fitness value f ;
- Step 5: Randomly choose two chromosomes from p , and apply the crossover operator to them;
- Step 6: Apply the mutation operator described in Section 3.3 to the new population;
- Step 7: Let the current population be the new population;
- Step 8: If the mean absolute error ϵ is less than 0.2%, then the convergence criterion is satisfied; stop. Otherwise, go to Step 4.

3.2. Coding and Initialization Methods

According to [16], the floating point representation is faster, more consistent and provides higher precision especially with large domains where binary coding would require prohibitively long representation. Compared with binary coding, the floating point coding has obvious advantages [17] as follows.

1. The floating point coding does not require a decoding process.
2. It represents a larger range of numbers in the genetic algorithm, which is convenient in larger spaces.
3. It also simplifies the traditional genetic algorithm and improves the computational efficiency.
4. Furthermore, the floating point encoding prevents the precision errors caused by the hexadecimal conversion.
5. It is easy to mix it with classical optimization methods.

Thus, the floating point representation is used in this paper. After the master's thesis dataset is preprocessed, the thesis number will participate in the coding operation.

Figure 3 shows the chromosome coding scheme in this paper. n is the total number of thesis to be sampled. p_{ij} ($1 \leq i \leq k$, $1 \leq j \leq m$) is the sequence number of the thesis taken from the thesis sample population randomly to generate the population in GA. Then, let $\omega_{ij} = p_{ij}/n$ to indicate the encoded value of the genes in chromosomes.

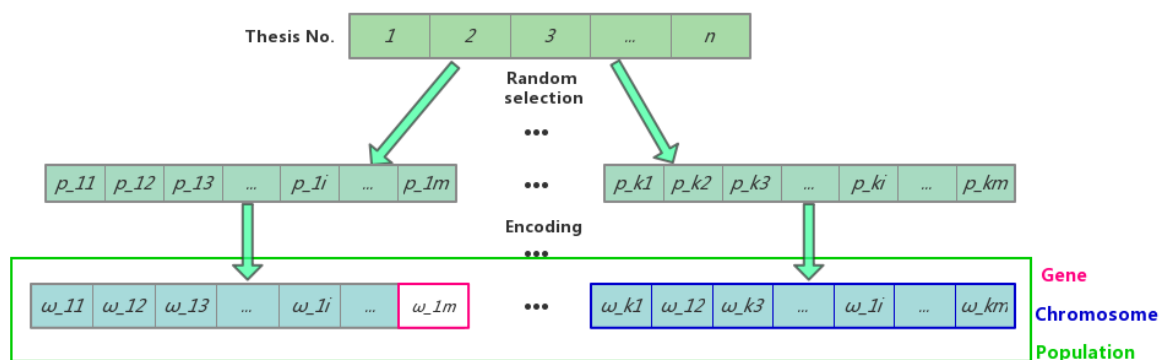


Figure 3. Initialization and encoding.

For example, for the first chromosome in Figure 3, assume n equals 10 and m equals 5: the theses with Serial Numbers 1, 3, 5, 7, 8 are selected. Accordingly, after encoding, the gene codes are 0.1, 0.3, 0.5, 0.7, 0.8.

3.3. Selection, Crossover and Mutation

The selection operator chooses some good individuals from the current population and transfers them to the next generation [18]. Methods used as the selection operator can be classified into two categories: proportional and elitist. The former category employs the probability of selection proportional to the fitness value of each individual. These methods allow maintaining a genetic diversity within the population of candidate solutions throughout the generations, which prevents GA from falling into local optima. However, these methods increase the time of convergence [19]. This paper uses the roulette wheel method making individuals with high-fitting values get a higher probability of survival. The basic idea under this method is that the probability of each individual being selected is proportional to its fitness value. The proportional selection method has enormous randomness and thus might be trapped in local extrema. The integration of simulated annealing factor into the genetic algorithm leads to avoid slow convergence rate and local extrema problem.

The crossover operator newly produces some better individual patterns [20], which is the main element of the search process in optimization. The encoding type used in GA is the major criteria for

selecting the crossover. The global convergence and search space must be considered for selecting the crossover operators. the effect of crossover operators in GA is dependent on the application, as well as encoding [21]. Here, the two-point arithmetic crossover is used.

$$\omega_1(t) = (\omega_{11}, \omega_{21}, \dots, \omega_{n1}) \quad (4)$$

$$\omega_2(t) = (\omega_{12}, \omega_{22}, \dots, \omega_{n2}) \quad (5)$$

where $\omega_1(t)$ and $\omega_2(t)$ are two chromosomes at time t . Let two random numbers generated at time t be $i(1 \leq i \leq n)$ and $j(0 < j < 1)$. Then, the two-point arithmetic crossover at the i -th locus of the above two chromosomes, respectively, as $\omega'_1(t)$ and $\omega'_2(t)$, is defined as follows:

$$\omega'_{i1} = j * \omega_{i2} + (1 - j) * \omega_{i1} \quad (6)$$

$$\omega'_{i2} = j * \omega_{i1} + (1 - j) * \omega_{i2} \quad (7)$$

The following descendants are propagated:

$$\omega_1(t+1) = (\omega_{11}, \omega_{21}, \dots, \omega'_{i1}, \dots, \omega_{n1}) \quad (8)$$

$$\omega_2(t+1) = (\omega_{12}, \omega_{22}, \dots, \omega'_{i2}, \dots, \omega_{n2}) \quad (9)$$

where, i is the locus involved in the crossover operation on the chromosome. n is the length of the chromosomes, that is, the number of genes. In this paper, it is the number of selected theses.

The mutation operator is employed to avoid establishing a uniform population unable to evolve [22]. According to [23], the NP-complete problems are intractable, i.e., they can be solved theoretically, but in practice, they take too long for their solutions to be useful.

Once a new generation is created after the crossover operation, the genetic process is performed iteratively until an optimal result is found. We have used the traditional genetic algorithm to solve the thesis sampling problem. More than 30,000 iterations were carried out, which took nearly 3.5 h, and the sampling effect was not ideal. By analyzing the traditional genetic algorithm, we found that the whole algorithm was time consuming. Since the mutation operator is directionless, we believe that some modifications could be made to the mutation operator based on the traditional genetic algorithm to speed up the convergence of the algorithm. Thus, an improved mutation strategy is introduced as Algorithm 1.

Figure 4, provided below, is for easier understanding of the proposed mutation operation.

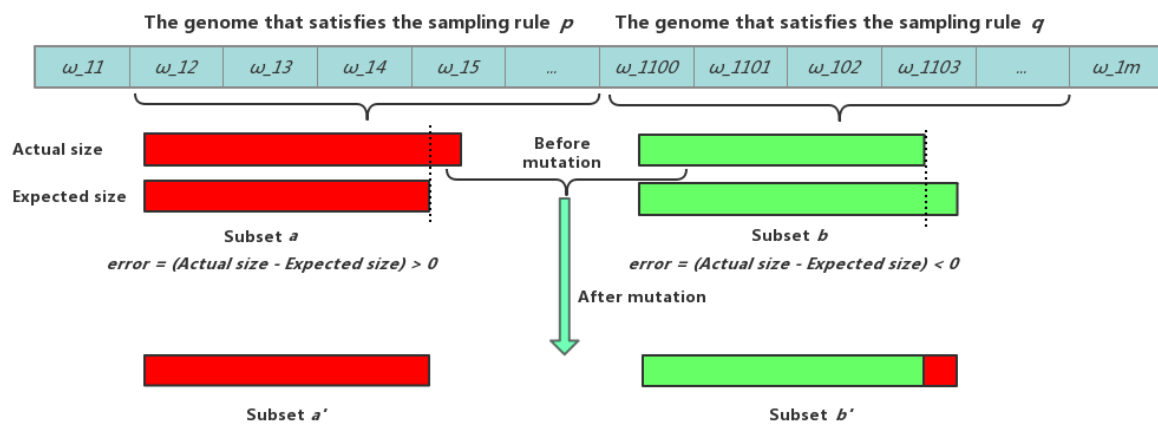


Figure 4. Schematic diagram of the improved mutation operation.

Algorithm 1: Mutation strategy.

```

1. Input: Thesis data, mutation probability, chromosome length, number of chromosomes and the fitness
2.   value of each chromosome.
3. Output: Mutated chromosomes
4. Initialize parameter:  $\epsilon$ 
5. for  $i = 1$  to the size of the population
6.   do
7.     for  $k = 1$  to the length of the chromosome
8.       do
9.         Denote the extracted thesis set (chromosome) as  $S$ , then calculate the error of the unique subset
10.        of  $S$  determined by the samplings rules according to the value of the genes.
11.       end for
12.       Assume that the subset  $a$  of set  $S$  has the maximum error, and subset  $b$  of set  $S$  has the
13.       minimum error.
14.       On the current chromosome, a mutation operation is performed by randomly replacing a gene
15.       corresponding to an element of set  $a$  with a gene that satisfies set  $b$ .
16.       Check whether the chromosome after the mutation meets the requirement that the same two genes
17.       do not exist on the chromosome.
18.       if the mutated chromosome meets the requirements
19.         The mutated chromosomes are accepted according to Equation (3);
20.       else
21.         Undo the mutation operation.
22.       end if
23.     end for
24.   end for

```

3.4. The Definition of Fitness Function

According to the sampling rules introduced in Section 2.2, the fitness function is defined as follows:

$$f = 1 - \frac{1}{n} \sum_{i=1}^n \sqrt{(P_i - E_i)^2}, \quad (10)$$

it is equivalent to:

$$f = 1 - \epsilon, \quad (11)$$

where n denotes the number of constrained conditions (sampling rules) and P_i and E_i indicate, respectively, the actual sampling rate and the expected sampling rate of the subset that satisfies the i -th sampling rule. ϵ is the mean absolute error. Then, the objective function of our proposed algorithm is,

$$g = \min\left(\frac{1}{n} \sum_{i=1}^n \sqrt{(P_i - E_i)^2}\right), \quad (12)$$

which is equivalent to the following equation,

$$h = \max\left(1 - \frac{1}{n} \sum_{i=1}^n \sqrt{(P_i - E_i)^2}\right). \quad (13)$$

4. Experimental Setup

This section describes the experimental setup for evaluating the proposed algorithm. The parameters used for the algorithm are also provided.

4.1. Experimental Setup

The improved GA was implemented in MATLAB and evaluated in the task for the sampling of Shanghai papers. For a fair comparison, the traditional GA and other comparison algorithms were also implemented in MATLAB. The evaluations were executed on a MacBook-Pro with 8 GB memory on a 2.7-GHz Intel Core i5 processor. The processed information of Shanghai municipal master's degree theses submitted from Shanghai in 2014 was used as input data. In total, 39,779 theses were used for the evaluation. The number of sampled theses and error rates were analyzed and provided in Section 5.

4.2. Parameters

Figures 5–7 show the analysis according to varying parameters. Based on the analysis and other information, the parameters were set as Table 1.

Table 1. Experiment parameters.

Size of initial population is 10
Size of population $sizepop = 4$
Probability of crossover $pcross = 0.97$
Probability of mutation $pmutation = 0.3$
Mode of selection $fselect = 'roulette'$
Method of coding $fcode = 'float'$
Mode of crossover $fcross = 'float'$
Temperature of annealing $T_0 = 100$
Temperature reduction parameter $k = 0.98$

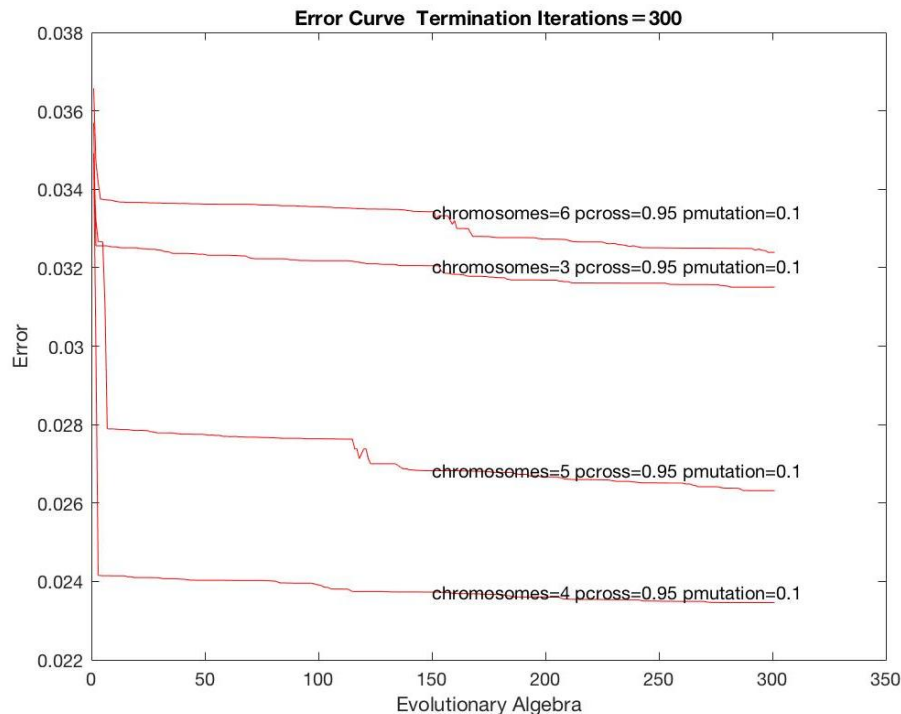


Figure 5. Effect of chromosome number on the objective function value.

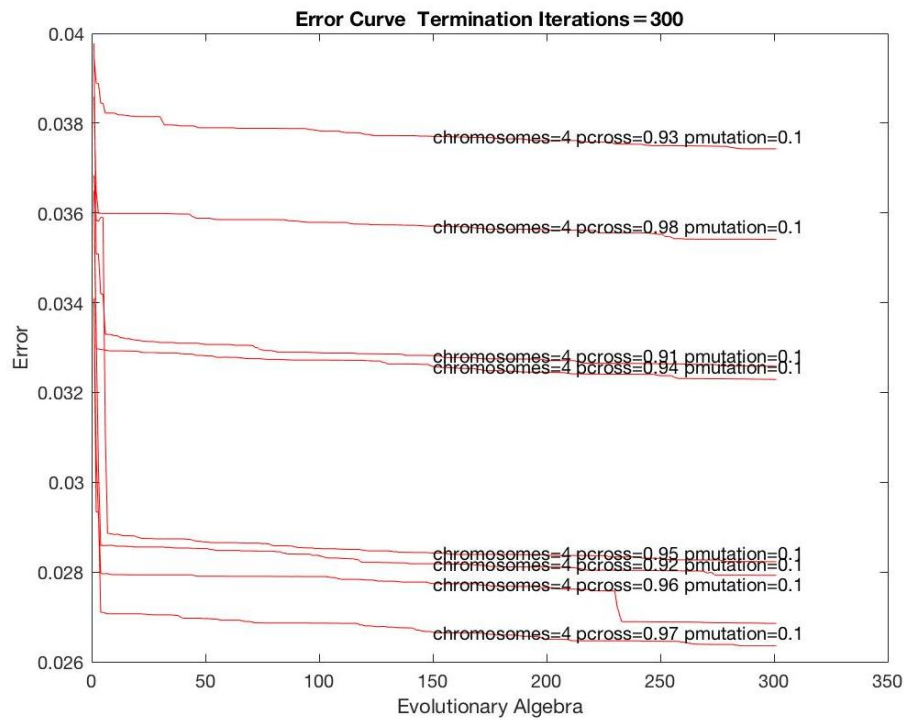


Figure 6. Effect of crossover probabilities on the objective function value.

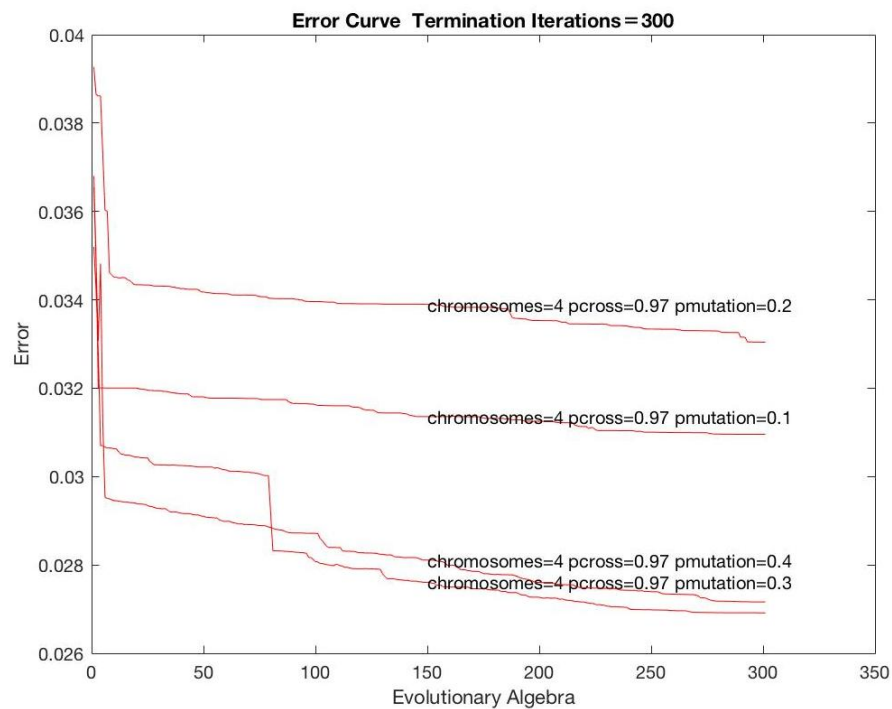


Figure 7. Effect of mutation probabilities on the objective function value.

4.3. Redundancy Reduction in Data

The source data shown as Figure 8a have too much redundant information relative to the data required for the experiment, so they need to be processed. The process mainly includes: adding missing data items, numbering the theses from 1 to 39,779 and adding attributes related to the sampling rules. To remove irrelevant data items, the processed data are as shown below.

[illegible]

	A	B	C	D	E	F	G	H	I	J	
1	id	name	nation	ethnicity	origin	major	IDCard	school	tutor	category	
2		1 Chen Qiulei	156	01	32	100201	32068319	10246		0	
3		2 Feng Yurong	156	01	37	100206	37232319	10246	2	0	
4		3 PENG WANGJING	158	97	71	100206	02315748	10246	3	0	
5		4 Zheng Yujia	156	01	23	100207	23020619	10246	2	0	
6		5 Wang Jing	156	01	32	100209	32060219	10246	3	0	
7		6 Liu Ninghua	156	01	31	100210	43525219	10246	1	0	
8		7 Xi Meili	156	01	42	100211	37292619	10246	1	0	
9		8 Zhang Li	156	01	41	100211	41122419	10246	1	0	
10		9 Gu Tingting	156	01	31	100214	31010519	10246	1	0	
11		10 Sun Huizhen	156	01	37	100214	37132119	10246	2	0	
12		11 Tan Cong	156	01	35	100214	50023419	10246	1	0	
13		12 Wang Chen	156	01	37	100214	37083119	10246	3	0	
14		13 Xu Han	156	01	34	100214	34262219	10246	2	0	
15		14 Li Jing	156	01	37	100601	37110219	10246	2	0	
16		15 Wang Zheng	156	01	37	100706	37142219	10246	4	0	
17		16 Wang Fangzi	156	01	31	100202	51302219	10246	15	0	
18		17 Wang Congmin	156	01	32	000000	32032519	10246	19	0	
19		18 YAMAKAWA YUKIKO	392		-1	-1	K5239400	10246	4	0	
20		19 LEE JEUNG HO	410		-1	-1	000000	M5128876	10246	4	0
21		20 PARK MINHEE	410		-1	-1	000000	M6739955	10246	1	0
22		21 MITA YUKIKO	392		-1	-1	000000	TK735243	10246	2	0
23		22 Zhai Changwen	156	01	37	100104	37088119	10246	1	0	
24		23 Zhao Yichao	156	01	31	100701	14020219	10246	2	0	
25		24 Xu Lingli	156	01	32	000000	32048219	10246	2	0	
26		25 Zhang Ting	156	01	32	071007	32028119	10246	5	0	
27		26 Guo Xiuping	156	01	44	070305	37142719	10246	4	0	
28		27 Liu Huancian	156	01	37	0702	37131119	10246	1	0	
29		28 Jin Zhigang	156	01	31	050301	33901119	10246	4	0	
30		29 Chan Wai Pan	446	97	82	020203	5151039	10246	6	0	

Figure 8. Examples of the experiment data.

Table 2 shows that the traditional genetic algorithm can generally meet the sampling requirements of Shanghai, but a significant drawback is that the sampling probability of some degree-granting units is zero. This is caused by the fact that these degree-granting units have fewer degree theses and traditional GA has considerable instability. After combining the simulated annealing algorithm, this

situation can be improved and then mixed with the new mutation strategy; the results were satisfactory and met the requirements of Shanghai thesis sampling.

Table 2. Sampling results under each algorithm.

Index	P_e	GA	H_a	$H_{a,m}$
1	10%	10.044%	10.044%	10.156%
2	10%	8.917%	9.251%	10.022%
3	5%	5.117%	5.015%	5.015%
4	5%	4.921%	5.042%	4.991%
5	5%	4.960%	5.035%	5.017%
6	5%	5.078%	5.000%	5.078%
7	5%	4.990%	4.991%	5.053%
8	5%	5.103%	5.029%	5.029%
9	5%	4.979%	5.025%	5.026%
10	5%	5.109%	5.109%	5.109%
11	5%	4.762%	4.762%	4.762%
12	5%	4.962%	4.962%	4.962%
13	5%	5.000%	5.000%	5.000%
14	5%	4.978%	5.004%	5.004%
15	5%	5.000%	5.000%	5.000%
16	5%	5.168%	5.048%	5.048%
17	5%	4.991%	4.991%	5.035%
18	5%	4.598%	4.885%	4.885%
19	5%	4.938%	5.010%	5.085%
20	5%	4.607%	4.878%	4.878%
21	5%	5.556%	4.938%	4.938%
22	5%	4.615%	5.384%	5.384%
23	5%	4.947%	5.015%	4.981%
24	5%	4.167%	5.000%	5.000%
25	5%	0.000%	2.272%	4.545%
26	5%	4.494%	4.494%	5.617%
27	5%	0.000%	4.764%	5.000%
28	5%	0.000%	5.274%	5.274%
29	5%	4.286%	4.286%	4.286%
30	5%	0.000%	0.000%	5.166%
31	5%	0.000%	5.882%	5.142%
32	5%	4.938%	4.761%	5.000%
33	5%	5.714%	5.263%	3.947%
34	5%	0.000%	0.000%	5.769%
35	5%	0.000%	0.000%	5.111%
36	5%	4.968%	5.000%	5.000%
37	5%	5.882%	5.882%	5.882%

In Table 3, GA is the traditional genetic algorithm, H_a is the hybrid approach with the annealing algorithm and $H_{a,m}$ is the hybrid approach with the annealing algorithm and the proposed mutation method.

Table 3. Comparisons on iterations, time cost and mean absolute errors.

	GA	H_a	$H_{a,m}$
Iterations	31,568	32,594	287
Time (min)	324.37	341.64	2.96
Mean absolute error (%)	1.155	0.644	0.164

In Table 3, the proposed method is compared to the traditional GA in time complexity and mean absolute errors. The analysis shows that adopting the annealing algorithm led to a locally optimal solution, but not to the globally optimal solution. The time consumption was still high, which might

lead to a long time to solve the problem, especially for the huge amount of data. Comparing the two methods, the proposed hybrid approach $H_{a,m}$ was greatly improved with the new mutation strategy regarding time and errors.

We can analyze the convergence characteristics of these algorithms according to Figure 9. Here, we assume that when the absolute error of a sampling index is less than 1%, then we can say that the convergence criterion of this sampling index is satisfied. If the mean absolute error of all these sampling indices is less than 0.2%, it satisfies the convergence criterion of the whole algorithm well.

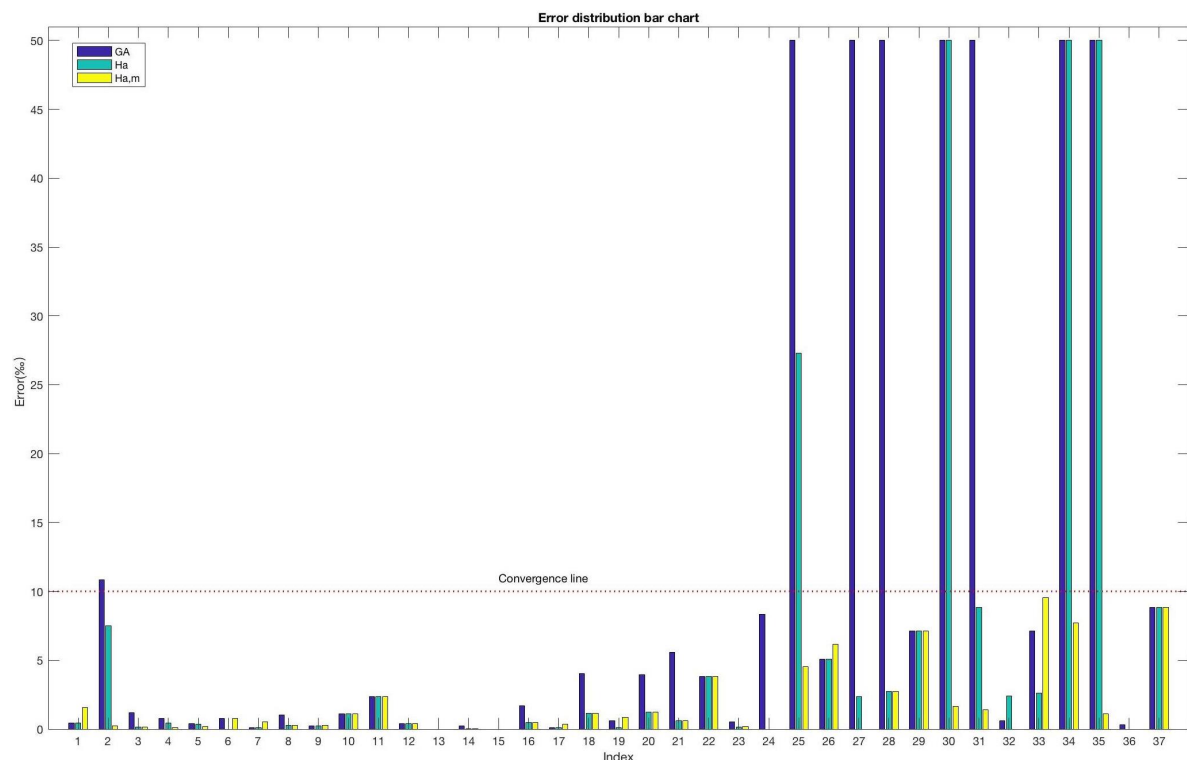


Figure 9. Error bar chart.

6. Conclusions and Future Work

The sampling of the master's degree thesis in Shanghai municipality is formulated as a problem of permutation and combination. The number of combinations of n theses taken m at a time is denoted by C_n^m . In this paper, n equals 39,779 and m equals about 2000. The number of combinations is too large, and meanwhile, it needs to satisfy three sampling rules. To address this problem, we proposed to combine the genetic algorithm and simulated annealing. The experimental results showed that the proposed algorithm outperformed the conventional genetic algorithm, providing hundreds of times lower complexity and satisfactory results. Furthermore, the proposed method satisfied the sampling requirements of Shanghai. In the future, we will expand the proposed algorithm to cope with the dynamic development of the demand for sampling in Shanghai. For example, sampling rules can also make relevant requirements for the attributes *ethnicity* and *origin* of the master's thesis.

Furthermore, since the initial parameters are obtained by experiments roughly and they are fixed, both good individuals and inferior individuals undergo the same probability of crossover and mutation operations. This can cause two very serious problems:

- (1) With the same probability, it can be said to be unfair, because, for good individuals, we should reduce the probability of cross mutation so that it can be preserved as much as possible; and for inferior individuals, we should increase the probability of crossover and mutation so that the inferior condition can be changed as much as possible.

- (2) The same probability cannot meet the needs of the evolution process of the population. For example, in the early iteration, the population needs a higher crossover and mutation probability, which has reached the goal of quickly finding the optimal solution. In the later stage of convergence, the population needs to be smaller. The crossover and mutation probability can help the population converge quickly after finding the optimal solution.

Therefore, the constant cross mutation probability affects the efficiency of the algorithm. Then, compared to traditional genetic algorithms, an adaptive genetic algorithm (AGA) is more suitable to solve this kind of multi-constraint problem.

Author Contributions: Conceptualization, S.J. Data curation, J.H. and Y.L. Formal analysis, X.W. Funding acquisition, S.J. Investigation, L.C. Methodology, S.J. Project administration, L.C. and X.W. Resources, S.J. Software, J.H. Supervision, S.J. Validation, X.W. Visualization, L.C. Writing, original draft, J.H. and Y.L. Writing, review and editing, J.H. and Y.L.

Funding: The theory of this research work is supported by the National Natural Science Fund Project (40976108; 61303097), the Shanghai Science Fund Project (17ZR1428400), the Shanghai Construction of Key Disciplines Fund Project (J50103), the Second (2016) Shanghai Research Project for Private University (2016-SHNGE-08ZD), the Shanghai University Graduate Innovation Fund Project (SHUCX070037; SHUCX120105) and the Provincial Government Education Comprehensive Reform of the National Education Reform Pilot Project (No. 10-109-091).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. State Council Degrees Committee, M.o.E. *Notice of the Ministry of Education on Printing and Distributing the Measures for the Examination of Doctoral Dissertations of Master Degree*; Bulletin of the Ministry of Education of the People's Republic of China: Beijing, China, 2014.
2. Xia, J.; Yang, X.; Shu, J. The concept of the construction of education informationization platform for postgraduate students from the provincial level and postgraduate program in Shanghai. *Degree Postgrad. Educ.* **2014**, *11*, 33–39.
3. Wu, C.W.; Aslam, M.; Jun, C.H. Variables sampling inspection scheme for resubmitted lots based on the process capability index Cpk. *Eur. J. Oper. Res.* **2012**, *217*, 560–566. [[CrossRef](#)]
4. Wang, Q. *New Progress in Shanghai Graduate Education: Commemorating the 30th Anniversary of Graduate Education*; Wang, Q., Ed.; Shanghai People's Publishing House: Shanghai, China, 2009.
5. Gao, L. Research on the Sampling System of Dissertation in China. Master's Thesis, Xiangtan University, Xiangtan, China, 2011.
6. Zhang, L. Research on Jiangsu province graduate dissertation evaluation mechanism. *Shanghai Educ. Eval. Res.* **2014**, *2*, 62–66.
7. Braun, H. On solving travelling salesman problems by genetic algorithms. In Proceedings of the International Conference on Parallel Problem Solving from Nature, Dortmund, Germany, 1–3 October 1990; pp. 129–133.
8. Deng, Y.; Liu, Y.; Zhou, D. An improved genetic algorithm with initial population strategy for symmetric TSP. *Math. Probl. Eng.* **2015**, *2015*. [[CrossRef](#)]
9. McCall, J. Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* **2005**, *184*, 205–222. [[CrossRef](#)]
10. Goldberg, D.E. *Genetic Algorithms in Search, Optimization and Machine Learning*; Addison-Wesley Co.: Boston, MA, USA, 1989; pp. 2104–2116.
11. Van Laarhoven, P.J.; Aarts, E.H. Simulated annealing. In *Simulated Annealing: Theory and Applications*; Springer: Heidelberg/Berlin, Germany, 1987; pp. 7–15.
12. Yu, H.; Fang, H.; Yao, P.; Yuan, Y. A combined genetic algorithm/simulated annealing algorithm for large scale system energy integration. *Comput. Chem. Eng.* **2000**, *24*, 2023–2035. [[CrossRef](#)]
13. Jha, S.; Menon, V. BbmTTP: Beat-based parallel simulated annealing algorithm on GPGPUs for the mirrored traveling tournament problem. In Proceedings of the High Performance Computing Symposium, Society for Computer Simulation International, Tampa, FL, USA, 13–16 April 2014; p. 3.
14. Kabova, E.A.; Cole, J.C.; Korb, O.; López-Ibáñez, M.; Williams, A.C.; Shankland, K. Improved performance of crystal structure solution from powder diffraction data through parameter tuning of a simulated annealing algorithm. *J. Appl. Crystallogr.* **2017**, *50*, 1411–1420. [[CrossRef](#)]

15. Assad, A.; Deep, K. A Hybrid Harmony search and Simulated Annealing algorithm for continuous optimization. *Inf. Sci.* **2018**, *450*, 246–266. [[CrossRef](#)]
16. Janikow, C.Z.; Michalewicz, Z. An Experimental Comparison of Binary and Floating Point Representations in Genetic Algorithms. In Proceedings of the Fourth International Conference on Genetic Algorithms, San Diego, California, USA, 1991; pp. 31–36.
17. Chen, L. Real coded genetic algorithm optimization of long term reservoir operation 1. *J. Am. Water Resour. Assoc.* **2003**, *39*, 1157–1165. [[CrossRef](#)]
18. Goldberg, D.E.; Deb, K. A comparative analysis of selection schemes used in genetic algorithms. In *Foundations of Genetic Algorithms*; Elsevier: New York, NY, USA, 1991; Volume 1, pp. 69–93.
19. Jebari, K.; Madiafi, M. Selection methods for genetic algorithms. *Int. J. Emerg. Sci.* **2013**, *3*, 333–344.
20. Chen, Y.; Yang, S.; Nie, Z. The application of a modified differential evolution strategy to some array pattern synthesis problems. *IEEE Trans. Antennas Propag.* **2008**, *56*, 1919–1927. [[CrossRef](#)]
21. Umbarkar, A.; Sheth, P. Crossover operators in genetic algorithms: A review. *ICTACT J. Soft Comput.* **2015**, *6*. [[CrossRef](#)]
22. Abdoun, O.; Abouchabaka, J.; Tajani, C. Analyzing the performance of mutation operators to solve the travelling salesman problem. *arXiv* **2012**, arXiv:1203.3099.
23. Sarkar, S.; Sinha, P.; Changder, N.; Dutta, A. Coalition Structure Formation using Parallel Dynamic Programming. In Proceedings of the 10th International Conference on Agents and Artificial Intelligence, Madeira, Portugal, 16–18 January 2018.



© 2018 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).