Determining an optimal subdivision of gene transfer partitions

MÁRK FARKAS
Department of Telecommunications and Media Informatics
Budapest University of Technology and Economics
Műegyetem rkp. 3-9. H-1111 Budapest
HUNGARY
farkas@tmit.bme.hu http://www.tmit.bme.hu

PÉTER FÖLDESI
Department of Logistics and Forwarding
Széchenyi István University
Egyetem tér 1. H-9026 Győr
HUNGARY
foldesi@sze.hu http://logisztika.sze.hu/

JÁNOS BOTZHEIM
Department of Automation
Széchenyi István University
Egyetem tér 1. H-9026 Győr
HUNGARY
botzheim@sze.hu http://www.sze.hu/~botzheim/

LÁSZLÓ T. KÓCZY
Faculty of Engineering Sciences
Széchenyi István University
Egyetem tér 1. H-9026 Győr
HUNGARY
koczy@sze.hu http://www.mtk.sze.hu

Abstract: - Bacterial memetic algorithms are widely used on discrete combinatorial problems, which are essential in the field of logistics and forwarding, such as the well known Traveling Salesman Problem. The original Bacterial Evolutionary Algorithm proposed by Nawa and Furuhashi [5] has a predefined set of operators such as bacterial mutation and gene transfer also known as infection. The traditional bacterial infection operator is proven to be far from optimal. The authors suggest an alternative gene transfer operator that is applied on the metric Traveling Salesman Problem [9]. This alternative infection algorithm has superior rate of convergence while reducing the risk of getting stuck in a local optima.

Key-Words: - Gene transfer, infection strategy, bacterial memetic algorithms, Traveling Salesman Problem

1 Introduction
Logistics and forwarding have always been important parts of human civilization. While currently the economy is shaken by a recession there is no reason to believe that the importance of logistics is following this declining tendency when observed on a long term. The most relevant aspect of this interdisciplinary science is cutting overall costs in both terms of monetary value and time.

In order to reach this goal one must formulate logistical problems on a solid mathematical foundation and
develop algorithms that are capable to solve real-life logistical problems on this abstract model. This principle often leads to tasks that are to be solved on a very complex problem domain. Many of them are encompassed by a well known complexity class, namely the class of NP-complete problems. Although these problems are by no means inextricable they still pose a formidable computational challenge when real-life input sizes are regarded. It is yet unknown whether these problems are computationally feasible on present-day computer architectures but it is widely believed that traditional algorithms in the sense of executable methods on an abstract deterministic Turing-machine may not be able to handle them efficiently, within polynomial time of the input size.

Theoretical science differs greatly from engineering and real-life economic applications as in our case a rough albeit still considerably good solution is often enough for our purposes. These facts lead to the development of deterministic approximation algorithms and non-deterministic stochastic heuristics and metaheuristics. These methods often find acceptable solution for a certain optimization problem. Such a family of metaheuristics forms the evolutionary algorithms, which is a population-based stochastic metaheuristics. These algorithms share several common properties that are to be described later. In our previous papers we found that a subset of evolutionary algorithms, the bacterial memetic algorithms are very efficient on solving discrete combinatorial optimization problems, such as the ones often found in logistics and forwarding.

Bacterial evolutionary algorithms were first suggested by Nawa and Furuhashi [5]. The algorithm has a very loose conformation as only the concept of the fundamental operators and control structures are defined. Most of the operators are defined in accordance with the optimization problem itself. However control structure properties are predetermined. We suggest a modification over the control structure of the bacterial infection operator that seems to improve the overall efficiency of the metaheuristics. The validity of this statement is verified in metric instances of a well-known problem in the field of logistics, the Traveling Salesman Problem.

2 Problem Formulation

In this section a brief overview is given on both the importance of gene transfer/infection operator in the control structure of the bacterial memetic algorithms and the example problem that on the efficiency of the operator, primarily the rate of convergence and the probability of finding the global optimum, is investigated.

2.1 Bacterial Memetic Algorithm

Bacterial Memetic Algorithms [7] are evolutionary methods that utilize the concepts of Lamarckian evolution. This is also known as heritability of acquired characteristics or soft inheritance. The main difference compared to traditional Darwinian methods [1][3] such as the Bacterial Evolutionary Algorithms (BEA) is the introduction of the concepts of memes. Memes are traits of acquired characteristics or cultural knowledge in the case of Lamarckian evolution [4]. In an evolutionary computational sense it narrowly means the inclusion of a local search heuristics into an existing evolutionary algorithm. Therefore memetic algorithms are a very broad subclass of evolutionary methods. In a more restrictive sense further on memetic algorithm will always refer to bacterial memetic algorithms encompassing a problem dependent local search algorithm such as the 2-opt or 3-opt deterministic local search method [8][10] in the case of logistical problems such as TSP and VRP. Besides local search bacterial memetic algorithms shares the rest of their operators with the Darwinian bacterial evolutionary algorithms.

2.1.1 Bacterial mutation

Bacterial mutation is a fundamental building block of the bacterial evolutionary algorithms [5]. Its role is very similar to that of the mutation operator in the standard evolutionary algorithms (EA). According to Holland’s Schema Theorem [1][3], mutation operator alters individuals, or bacteria in the case of BEA, implicitly sampling multiple hyperplanes such that mutated individuals or bacteria are going to sample adjacent hyperplanes. As contrary to standard EAs, Bacterial Evolutionary Algorithms do not include selection. This is mainly because bacterial mutation always keeps the most fit clone bacteria after each consecutive segment-wise mutation. This can be regarded as selection on a certain extent. By utilizing a large number of clones during the course of bacterial mutation the whole process can be more effective than the standard mutation operator in the evolutionary algorithms. More about the course and importance of bacterial mutation can be found in [5] and [7].

2.1.2 Bacterial infection

The primary role of the bacterial infection is to share partial sub-optimal solutions across the population. This goal is achieved by passing genetic information from a donor bacterium into a recipient bacterium [5][7]. The means of information passing greatly depends on the underlying optimization problem and the representation of a solution instance in the problem domain as a bacterial chromosome. Despite that dependency on the problem, the control structure itself is somewhat strictly formulated. Initially
the population is ordered by its fitness and divided into two separate partitions according to its fitness value. Each infection takes two distinct bacteria from both the first and the second partition of the population. Those coming from the first half takes the role of information source or donor while those coming from the less fit second half takes the role of information destination or recipient.

While this strategy seems applicable there is no reason to believe that it is the most efficient one by any means. This disbelief leads us to the development of this modified gene transfer method.

2.2 Traveling Salesman Problem

Traveling Salesman Problem [9] is formally defined as a graph problem where the task is to find a Hamilton circle (a route containing all vertices at least and at most once). This is formally described as a series of distinct vertices in a G graph:

\[ G_{\text{TSP}} = (V_{\text{cities}}, E_{\text{conn}}) \]

\[ V_{\text{cities}} = \{v_1, v_2, \ldots, v_n\}, E_{\text{conn}} \subseteq \{(v_i, v_j) | i \neq j\} \]

\[ d: V_{\text{cities}} \times V_{\text{cities}} \rightarrow \mathbb{R}^+, C := [d(v_i, v_j)] \]

Equation 1: The Traveling Salesman Problem as a Graph Problem

An instance in the TSP problem is a set of interconnected cities and a distance matrix. The set of feasible solutions of an instance comprises of all possible permutations of the series of vertices while the measure is the sum of distances of the whole vertex series. The goal function is to minimize the directed Hamiltonian circle described by the permutation:

\[ I_{\text{TSP}} = (G_{\text{TSP}}, C), \forall x \in I_{\text{TSP}} : \]

\[ f_{\text{TSP}}(x) = (v_{x(1)}, v_{x(2)}, \ldots, v_{x|G_{\text{TSP}}|}) \]

\[ \forall x \in I_{\text{TSP}}, y \in f(x) : m_{\text{TSP}}(x, y) = \sum_{i=1}^{|G_{\text{TSP}}|} C_{v_{x(i)}, v_{x(i+1)}} \]

Equation 2: An Instance and Measure of a TSP Problem

Several subsets of the traveling salesman problem [11] may be interesting as optimal operator set and control structures may be different. In the followings however, only metric traveling salesman problem is investigated:

2.2.1 Metric TSP

Metric TSP is a subset of the generalized traveling salesman problem [11] where the only restriction is that the distance function shall be a metric over a vector space e.g. a Euclidean distance metric over the 2D Euclidean vector space. This can be expressed by defining the following two axioms on the distance function:

\[ d(x, y) = 0 \iff x = y \]

\[ d(x, z) \leq d(x, y) + d(y, z) \]

Equation 3: The Axioms of a Metric

The second axiom is also known as the triangle inequality. This axiom has a great importance in the case of the metric Traveling Salesperson Problem, as the shortest tour containing all the cities at least once cannot contain any city more than once.

2.3 A modified infection strategy

When infection or gene transfer operator is regarded the original bacterial evolutionary algorithm [5] sorts the entire population after mutation and divides it into two subgroups with an equal or almost equal size (depends on the cardinality of the set). Consecutive source and destination bacteria are selected from these respective sets. After performing an infection on a destination bacterium it is reinserted into the ordered population and population is divided into two subgroups again.

An alternative technique would be to create the source group from the k% percent of most fit bacteria and place the rest of them into the group containing the recipient bacteria. This technique is assumed to behave differently in terms of rate of convergence in some cases. One has to note though that the rate of convergence depends greatly on the proper estimation of the value of k.

First and foremost, measures are made in order to determine the optimal k value in each test case. Secondly a heuristics is suggested that may be able to predict the optimal k value before evaluating the algorithm.

3 Problem Solution

It is assumed that using this modified infection strategy may alter the divergence of the population. When the divergence of a population falls below a certain level the population is endangered by a premature convergence. One way to measure the divergence of the population is to measure its standard deviation across the population and the standard deviation of the bacteria possessing the maximal fitness in the given population across the entire course of simulation. These attributes are to be measured on reference TSP test cases. Measurements are to be made by increasing the value of k by the finest possible grain. Each test is conducted at least 10 times and the average of them is chosen. Thus 5% resolution means over 100 measurements on a single test instance.

Measurements were made with the LogFerm v.1.0 RC2 logistics framework [11], which is available on request from the authors.
Parameters were adjusted according to previous experiences; these parameters produced overall good results and were able to result in the global optimum from 50 out of 50 times in TSP test cases containing less than, or equal to 100 cities. This factor is emphasized because of the very stochastic nature of the algorithm.

- Number of clones: 10
- Mutation segment length: 7
- Loose segment possibility: 0.5
- Number of infections: 50
- Infection segment length: 15
- % of 2-opt local search: 30%
- % of 3-opt local search: 10%
- Eugenic mutation: ON
- Initial tour construction: SNN

3.1 The Berlin 52 test case

The Berlin 52 [14] is a small test case having 52 cities and the shortest Hamiltonian tour containing each city at least and at most once has a total length of 7544.36 in our case (the actual results in the reference TSPLIB repository are shorter due to rounding error in the Concorde implementation).

The rate of convergence can be interpreted as a nonlinear function of k. In this case the function has two local minima at 40% and 5%. It should be noted that the local minima at 5% has still got a superior rate of convergence over the interval between 35% and 50%. The function has 3 local maxima at 1%, 10% and 45% respectively, although the maximum at 45% falls below the other 2 maxima by almost 10 percent. The function between the maximum point at 10% and the minimum point at 40% acts almost linearly decreasing. This indicates that the normal infection strategy is far from optimal and there is headroom for improvement by over ten percent. This does not sound too much at the first glance. This is mainly due to the fact that more than 100 generations were used in our test case and in almost all scenarios the algorithm converged after 50 successive generations. If one takes a look on the average number of consecutive generations required to achieve the global optimum the picture changes dramatically however. In this example that means using an optimal k=10% population subdivision the algorithm converges after 20 generations on average, while using a worst-case k=40% population subdivision the algorithm usually converges after only 38 successive generations. It is therefore crucial to select the value of k optimally as the optimal rate of convergence might be more than twice the worst convergence speed achieved. If the comparison is made between the optimal and the default infection strategy then the optimal strategy outperforms the default infection strategy; which is the 50-50% population subdivision, by 50%. From this point of view determining an optimal population subdivision is definitely worth a while. All setups were able to reach the global optimum in every conducted test run; therefore the question is reduced solely to finding the optimal k value having the greatest rate of convergence.

The table below contains numeric results on the average rate of convergence; where 1 means the algorithm has achieved global optimum after tour construction and 0 means it has never managed to reach the global optimum at all. Standard deviation values contain the average of the minimal standard deviations in the population according to the respective fitness value of the bacteria across the entire generation. For the sake of convenience convergence rates are presented as a graph:

<table>
<thead>
<tr>
<th>k%</th>
<th>Std. Dev.</th>
<th>Conv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>4056.03</td>
<td>0.799</td>
</tr>
<tr>
<td>5%</td>
<td>4258.06</td>
<td>0.726</td>
</tr>
<tr>
<td>10%</td>
<td>4117</td>
<td>0.798</td>
</tr>
<tr>
<td>15%</td>
<td>4149.2</td>
<td>0.779</td>
</tr>
<tr>
<td>20%</td>
<td>4160.2</td>
<td>0.776</td>
</tr>
<tr>
<td>25%</td>
<td>4261.6</td>
<td>0.736</td>
</tr>
<tr>
<td>30%</td>
<td>4232.4</td>
<td>0.737</td>
</tr>
<tr>
<td>35%</td>
<td>4445.2</td>
<td>0.658</td>
</tr>
<tr>
<td>40%</td>
<td>4396.2</td>
<td>0.623</td>
</tr>
<tr>
<td>45%</td>
<td>4281.2</td>
<td>0.721</td>
</tr>
<tr>
<td>50%</td>
<td>4344.5</td>
<td>0.706</td>
</tr>
</tbody>
</table>

The Berlin52 –率为 converged instances

The Berlin52 – OPTIMAL TOUR INSTANCE

TABLE I

THE BERLIN52 TSP TEST CASE

BERLIN52 – RATE OF CONVERGENCE
3.2 The Smith-Thompson 70 test case

The Smith-Thompson 70 [15] is a medium test case having 72 cities and a shortest tour of 677.1 (it is actually a few units shorter in the previously mentioned TSPLIB reference repository, this is due to the fact that Concorde software rounds the distance between adjacent cities to 2 decimal places; whereas LogFerm calculates distances using 64-bit double precision arithmetic)

<table>
<thead>
<tr>
<th>k%</th>
<th>Conv.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>0.799</td>
</tr>
<tr>
<td>5%</td>
<td>0.726</td>
</tr>
<tr>
<td>10%</td>
<td>0.798</td>
</tr>
<tr>
<td>15%</td>
<td>0.779</td>
</tr>
<tr>
<td>20%</td>
<td>0.776</td>
</tr>
<tr>
<td>25%</td>
<td>0.736</td>
</tr>
<tr>
<td>30%</td>
<td>0.737</td>
</tr>
<tr>
<td>35%</td>
<td>0.658</td>
</tr>
<tr>
<td>40%</td>
<td>0.623</td>
</tr>
<tr>
<td>45%</td>
<td>0.721</td>
</tr>
<tr>
<td>50%</td>
<td>0.706</td>
</tr>
</tbody>
</table>

Similarly to the previously examined test case the table above contains the rate of convergence measured on over 100 successive generations. Standard deviation attribute is omitted here as it does not really provided additional information on the correspondence between the value of k and the absolute rate of convergence, as it had shown increasing tendency when the value of k was increased, regardless of the rate of convergence; thus it is not measured any more. The graph below depicts the relative speed of convergence as compared to the fastest converging population subdivision; the higher percentage means a higher speed of convergence:

The result has a very high resemblance to the previously observed result in the case of the berlin52 TSP instance. It has a global minimum point at exactly 40% of the population while the global maximum is around 5%. Between the global minimum and maximum; that is between 5 and 40 k values, the rate of convergence is decreasing.

The rate of convergence is more than 2 times faster in the case of k=5 where it has the highest value than the k=40%, which is the worst case scenario. If normal infection strategy and the k=5% optimal population subdivision is examined then one may find out that optimal population subdivision technique is almost twice as fast as the normal bacterial infection strategy. Another important aspect is that using the optimal population subdivision the algorithm has always managed to find the global optimum after 100 consecutive generations; while the standard infection strategy often required more than 100 (between 100 and 250) successive iteration.

If the rate of convergence is regarded as the function of k then both functions (berlin52 and st70) have very similar shapes, although the place of the global optimum does not exactly match. This is mainly due to the fact that in the case of the former test only 100 bacteria were used while in the case of the later, st70 case exactly thrice as many bacteria were being utilized. An observation is made that in the case of metric traveling salesman problems only a dozen bacteria are enough to be a donor while the rest of the bacteria can be recipients as this strategy leads to the highest possible rate of convergence.

3.3 Explaining the results

A possible explanation of this effect is the following: while the normal infection strategy provides a high level of divergence in the population [5][7]; the algorithm quickly reaches local optima, which has approximately similar fitness value to the global optimum. In most cases the fittest tour that represents the local optimum has large segments of suboptimal tours from the global optimum represented by the Hamiltonian circle [9] with minimum aggregate length. When normal infection strategy is utilized there is less chance to select this local optimum as a donor bacterium because the donor part of the subdivided population is considerably larger when
large k values are used. Therefore less optimal bacteria has higher chance of selection and thus a higher chance to infect recipient bacteria, which are otherwise contains only minor subsections of the optimal tours. In this case neither the donor nor the recipient bacteria are selected from the most optimal one, or to be more precise less optimal donor/recipient bacteria pairs have equal probability. As k increases so does the size of the donor bacteria set, which results in the decrease of the probability of selecting a near optimum sampling point represented by a bacterium having an above than average fitness.

On the other hand the size of the set containing recipient bacteria decreases and contains less valuable bacteria than in the case of smaller k values. By choosing the k smaller than in the case of traditional bacterial infection the chance of choosing a proper donor and recipient pair becomes higher. The overall divergence of the population is affected on a much less extent. When the value of k converges to 0 the divergence diminishes greatly which has a huge impact on the rate of convergence. This effect contributes to the declining rate of convergence between the global optimum at around k=5-10% and k=0%.

4 Conclusion

Extensive test runs conducted on small and medium sized metric traveling salesperson problems showed that traditional bacterial infection strategy is far from being optimal. An alternative infection strategy is suggested where an adjustable parameter (k) determines the percentage of donor/recipient population space partitioning. It is shown that using k=5-10% values have the highest rate of convergence amongst such algorithms and these parameters did not have any negative side effects on the algorithm itself; it was successful on reaching global optimum in all cases. A possible explanation was given to explain this behavior of the Bacterial Memetic Algorithms. Other alternative infection strategies and larger TSP instances may be a subject of further investigation.

References:
[13] Peter Turney, Myths and Legends of the Baldwin Effect, Institute for Information Technology, National Research Council, Canada
[14] M. Groetschel, 52 locations in Berlin
[15] Smith and Thompson, TSP instance containing 70 locations.