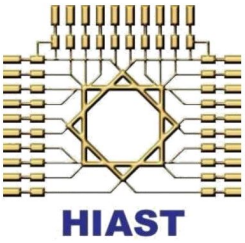
Syrian Arab Republic
Higher Institute for Applied Sciences and Technology
Informatics Department
Fourth Year



Using Genetic Algorithms to Find Approximations for the Minimum Vertex Cover Problem

Keywords: genetic algorithms, NP-complete, combinatorial optimization, non-deterministic algorithms, approximation algorithms, minimum vertex cover.

Author: *Farouk Hjabo*
Academic Supervisor: *Dr. Said Desouki*
General Supervisor: *Dr. Kadan Aljoumaa*
Langauge Supervisor: *Mr. Fahmi Alammareen*

February 23, 2018

Abstract

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

"It is not the strongest of the species that survives, nor the most intelligent, but the one most responsive to change."

— Charles Darwin

Contents

1 Introduction

Once the NP-hardness of a combinatorial optimization problem is established, the search for an optimal solution is abandoned. The goal then becomes one of finding a good heuristic, i.e. a polynomial running time algorithm that can find solutions close to the optimal. In most cases, traditional heuristics are problem dependent; a heuristic is tailored to the specific problem it is trying to solve.

In this paper, we present an alternative approach that uses genetic algorithms as a generalized heuristic for solving NP-hard combinatorial optimization problems. The application of a genetic algorithm is demonstrated here for the *minimum vertex cover* problem. These algorithms have been successfully applied to a broad range of problems. This wide range can be tackled by genetic algorithms mainly due to the fact that they work with an encoding of the domain rather than with the problem domain itself.

2 Genetic Algorithms

2.1 The Intuition Behind GAs

Genetic Algorithms (GAs) [3] are population based search algorithms, where by repeated use of genetic operations, such as ***mutation***, ***selection***, ***crossover***, etc... Successive new generations of better populations in the direction of search objectives are created. They are inspired by Darwinian principles based on natural evolution. In other words, the main idea behind genetic algorithms is that only the ***fittest*** individuals will survive.
Main advantage of genetic algorithms is that they ideally do not make any assumption about the underlying problem, hence they are suitable to tackle a wide range of diverse problems in engineering, art, biology, economics, marketing, genetics, operations research, robotics, social sciences, physics, politics, chemistry, etc...

So what is a GA? A typical GA consists of the following:

1. a number, or ***population***, of candidate solutions to the problem.
2. a way of calculating how good or bad the individual solutions within the population are. i.e. how ***fit*** an individual is?

3. a method for mixing fragments of the better solutions to form new, on average even better solutions. Which permits the population to ***evolve*** naturally.
4. a ***mutation*** operator to avoid permanent loss of diversity within the solutions. This allows introducing new information in the population.

2.2 Typical GA Schema

The basic iteration cycle of a genetic algorithm proceeds on a population of individuals, each of which represents a search point in the space of potential solutions of a given optimization problem. In case of a canonical genetic algorithm, each individual is a binary vector $\vec{x} = (x_1, \ldots, x_n) \in \{0,1\}^n$ of fixed length n. The fitness function $f : \{0,1\}^n \to \mathbb{R}$ provides a quality measure which is used by the selection procedure to direct the search towards regions of the search space where the average fitness of the population increases. The recombination operator allows for the exchange of information between different individuals, and mutation introduces innovation into the population.
This cycle is repeated until a termination criterion is fulfilled. In most cases, the algorithm is terminated after a certain number of iterations of the basic cycle or when a satisfactory fitness level has been reached. refer to figure 1 for a visual representation of this process.

Initialization The population size (denoted as u) depends on the nature of the problem, but typically contains several hundreds or thousands of possible solutions. Often, the initial population is generated randomly, allowing the entire range of possible solutions (the search space). Occasionally, the solutions may be 'seeded' in areas where optimal solutions are likely to be found.

Selection Normally, selection in genetic algorithms is a probabilistic operator which uses the relative fitness $p_s(\vec{x_i}) = f(\vec{x_i}) / \sum_{j=1}^{u} f(\vec{x_j})$ to serve as selection probabilities (u denotes the population size). This selection operator is called proportional selection. If the problem under consideration is a minimization one, or if the fitness function can take negative values, then $f(\vec{x_i})$ has to be linearly transformed before calculating selection probabilities. This technique known as linear dynamic scaling is commonly used in genetic algorithms (see [13], pp. 123-124, or [14]).

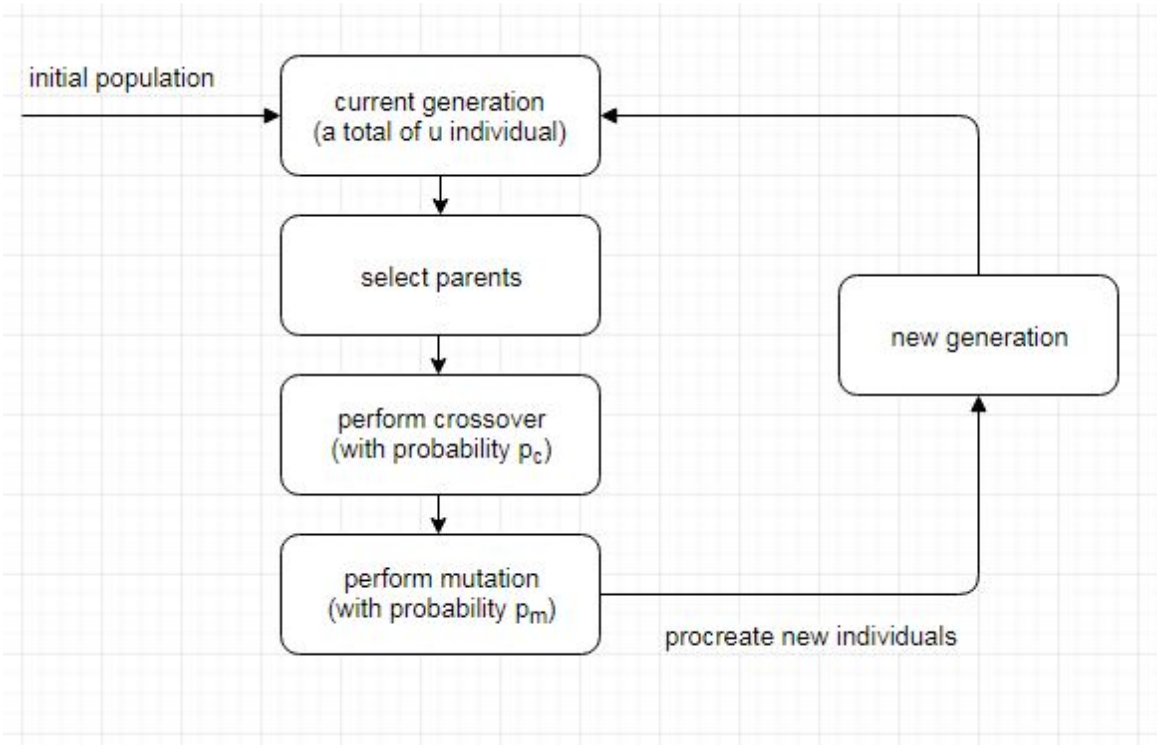


Figure 1: general schema for genetic algorithms

Crossover The recombination (crossover) operator allows for the exchange of information between different individuals. The original one-point crossover [9] works on two parent individuals (which are randomly chosen from the population, see the previous paragraph todo) by choosing a crossover point $\chi \in \{1, \ldots, n-1\}$ at random and exchanging all bits after the χ^{th} one between both individuals. The crossover rate p_c (e.g. $p_c \approx 0.6$) determines the probability to undergo crossover (see todo for more details). The crossover operator can be extended to a generalized multi-point crossover [10] or even to uniform crossover, where an it is randomly decided for each bit whether to exchange it or not [19]. The strong mixing effect introduced by uniform crossover is sometimes helpful to overcome local optima.

Mutation Innovation, i.e. new information, is introduced into the population by means of mutation, which works by inverting bits with a small probability p_m (e.g. $p_m \approx 0.001$) Though mutation is often interpreted as a rather unimportant operator in genetic algorithms [9], recent theoretical work gives strong evidence for an appropriate

choice of a mutation rate $p_m = 1/n$ on many problems [2; 12].

For better understanding of figure 1

Algorithm 1 Gentic Algorithm(n, u, p_c, p_m, f)

```
1: P ← {}                                              ▷ P contains the population
2: loop u times                                                  ▷ initialization
3:     add a randomly generated binary vector of length n to P
4: end loop
5: while termination criterion is not fulfilled do
6:     Q ← {}                                          ▷ Q will hold the new generation
7:     loop u/2 times
8:         select x, y from P using proportional selection              ▷ selection
9:         with probability p_c perform crossover between x, y          ▷ crossover
10:        for each bit in x flip it with probability p_m               ▷ mutation
11:        for each bit in y flip it with probability p_m               ▷ mutation
12:        add x, y to Q
13:    end loop
14:    P ← Q
15: end while
```

3 The Minimum Vertex Cover Problem

The problem of finding a minimum vertex cover is a classical optimization problem in computer science and is a typical example of an NP-hard optimization problem that has an approximation algorithm. The mvcp of an undirected graph $G = (V, E)$ where V is the set of vertices and E denotes the set of edges, consists in finding the smallest subset $V' \subseteq V$ such that $\forall \langle i, j \rangle \in E$, we have $i \in V'$ or $j \in V'$ (or both). V' is said to be a vertex cover of G. The following is a formal definition of the mvcp in which we make use of Stinson's terminology for combinatorial optimization problems [14]:

Problem instance: A graph $G = (V, E)$, where $V = \{1, 2, \ldots, n\}$ is the set of vertices and $E \subseteq V \times V$ the set of edges. An edge between vertices i, j is denoted by

the pair $\langle i, j\rangle \in E$. We define the adjacency matrix e_{ij} according to:

$$e_{ij} = \begin{cases} 1 & \text{if } \langle i, j\rangle \in E \\ 0 & \text{otherwise} \end{cases}$$

especially, we have $e_{ii} = 0$

Feasible solution: A set $V' \subseteq V$ such that $\forall \langle i, j\rangle \in E, i \in V' \vee j \in V'$

Objective function: The size $|V'|$ of the vertex cover V'.

Optimal solution: a vertex cover V' that minimizes $|V'|$.

an interesting property of the mvcp is that by solving it, we also have a solution for two other graph problems: the *maximum independent set problem*[1] and the *maximum clique problem*[2] The close relationship between these problems is characterized by the following lemma [6].

Lemma 1

For any graph $G = (V, E)$ and $V' \subseteq V$, the following statements are equivalent:

- *V' is the minimum vertex cover of G.*
- *$V - V'$ is the maximum independent set in G.*
- *$V - V'$ is the maximum clique in $\bar{G} = (V, \bar{E})$, where $\bar{E} = \{\langle i, j\rangle \in V \times V : \langle i, j\rangle \notin E\}$*

Consequently, one can obtain a solution of the maximum independent set problem by taking the complement of the solution to the minimum vertex cover problem. A solution to the maximum clique problem is obtained by taking the complement of the minimum vertex cover of $\bar{G}$.

[1] misp: given $G = (V, E)$, find the largest subset $V' \subseteq V$ such that $\forall i, j \in V', \langle i, j\rangle \notin E$

[2] mcp: given $G = (V, E)$, find the largest subset $V' \subseteq V$ such that $\forall i, j \in V', \langle i, j\rangle \in E$

3.1 Related Algorithms

Since we are interested in covering all the edges of the given graph by using as few nodes as possible, one might be tempted to use a greedy-based heuristic to tackle the mvcp. The algorithm consists in repeatedly selecting a vertex of highest degree (the node that covers as many of the remaining edges as possible), and removing all of its incident edges. This is not a good strategy as was demonstrated by Papadimitriou and Steiglitz ([10], p. 407).
They considered regular graphs, each of which consists of three levels. The first two levels have the same number of nodes while the third level has two nodes less than the number of nodes found on the previous two levels. More precisely, each graph consists of $n = 3k + 4 (k \geq 1)$ nodes; $k + 2$ nodes on the first level, labeled $1, 2, \ldots, k + 2$; followed by $k + 2$ nodes on the second level, labeled $k + 3, \ldots, 2k + 4$; while the k nodes of the third level are labeled $2k + 5, \ldots, 3k + 4$. The regular graph for $k = 4$ can be found in figure **??**. A minimum vertex cover is obtained by choosing all the nodes of the second level, but the greedy strategy would start with the nodes of the third level, since these have highest degree. Consequently, the greedy strategy finds a solution of size $2k + 2$ (since it has to select $k + 2$ nodes in addition to the nodes of the third level), while the optimal solution has size $k + 2$. However this algorithm is not totally useless. It may be shown that it always achieves the ratio $\Theta(\lg n)$, where n is the number of vertices. That is, if $C^*(n)$ is the size of the optimal solution and $C(n)$ is the size of the solution found by the aforementioned algorithm. Then we have $\frac{C(n)}{C^*(n)} = \Theta(\lg n)$ [9], p-323.

Fortunately, a simpler and better algorithm exists. The following simple algorithm is surprisingly the best approximation algorithm known for the mvcp ([9], p. 301).

Algorithm 2 vercov

1: $C \leftarrow \{\}$ ▷ C contains the vertex cover being constructed
2: $E' \leftarrow E$
3: **while** $E' \neq \phi$ **do**
4: randomly choose $\langle u, v \rangle \in E'$
5: $C \leftarrow C \cup \{u, v\}$
6: $E' \leftarrow E' - \{\langle x, y \rangle \in E' : x = u \vee y = v \vee x = v \vee y = u\}$
7: ▷ remove from E' every edge incident on either u or v
8: **end while**
9: **return** C

This is a *2-approximation algorithm*. Which means that the size of the vertex cover it returns is guaranteed to be no more than twice the size of an optimal vertex

cover ([1], p. 968). More precisely, if C^* is the size of the optimal solution and C is the size of the solution returned by Algorithm 2 then the inequality $C \leq 2C^*$ must hold (notice that here the approximation doesn't depend on the number of vertices n).

3.2 GA for the mvcp

In order to apply GA to solve the mvcp, a cover V' is represented by a binary vector $\vec{x} = x_1 x_2 \dots x_n$ where $x_i = 1$ if the i^{textth} node is in V', and $x_i = 0$ if it is not. Using this representation, genetic algorithm is applied to minimize the following fitness function: [0]

$$\begin{aligned} f(\vec{x}) &= \sum_{i=1}^{n} \left(x_i + n \cdot (1 - x_i) \cdot \sum_{j=i}^{n} (1 - x_j) e_{ij} \right) \\ &= \sum_{i=1}^{n} x_i + n \cdot \sum_{i=1}^{n} \sum_{j=i}^{n} (1 - x_i) \cdot (1 - x_j) \cdot e_{ij} \end{aligned} \tag{1}$$

The term $\sum_{i=1}^{n} x_i$ of $f(\vec{x})$ determines the size of the potential vertex cover represented by $\vec{x}$, while the term $n \cdot \sum_{i=1}^{n} \sum_{j=i}^{n} (1 - x_i) \cdot (1 - x_j) \cdot e_{ij}$ penalizes sets V' that are not covers by adding a penalty of magnitude n for each edge e_{ij} for which $i \notin V' \wedge j \notin V'$ (note that e_{ii} is previously defined to be zero). Consequently, any infeasible solution will have a fitness greater than or equal to n, and for any feasible solution the second term will drop to zero. This fitness function was developed according to the following design principles that are important for a successful penalty function approach [11 || 13 || 7]:

- Infeasible binary vectors are guaranteed to yield fitness values which are inferior to fitness values of the worst feasible solutions.
- The penalty should be graded, i.e., fitness values should improve as solutions approach feasible regions of the search space. In other words, the more edges a solution covers, the fitter it is. notice that this is guaranteed by the second term in equation 1, where for each uncovered edge the fitness is penalized with the value of n.

4 Experimental Results

Lorem [1] ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua. Ut enim ad minim veniam, quis nostrud exercitation ullamco laboris nisi ut aliquip ex ea commodo consequat. Duis aute irure dolor in reprehenderit in voluptate velit esse cillum dolore eu fugiat nulla pariatur. Excepteur sint occaecat cupidatat non proident, sunt in culpa qui officia deserunt mollit anim id est laborum.

5 Conclusion

sit amet, consectetur adipisicing elit, sed do eiusmod

References

[1] P. A. M. Dirac. *The Principles of Quantum Mechanics*. International series of monographs on physics. Clarendon Press, 1981.

[2] A. Einstein. "Zur Elektrodynamik bewegter Körper. (German) [On the electrodynamics of moving bodies]". In: *Annalen der Physik* 322.10 (1905), pp. 891–921.

[3] D. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Artificial Intelligence. Addison-Wesley Publishing Company, 1989.