The Hitting Set Problem and Evolutionary Algorithmic Techniques with ad-hoc Viruses (HEAT-V)

Vincenzo Cutello[1] and Francesco Pappalardo[1]

Dept. of Mathematics and Computer Science, University of Catania
V. le A. Doria, 6-I, 95125 Catania, Italy
{cutello,francesco}dmi.unict.it

Introduction. The *Weighted Minimum Hitting Set Problem (WMHSP)* and the standard *Minimum Hitting Set Problem (MHSP)*, are combinatorial problems of great interest for many applications. Although, these problems lend themselves quite naturally to an evolutionary approach. To our knowledge, there are no significant results of evolutionary algorithms applied to either the WMHSP or the MHSP, except for the results contained in Cutello et al., (2002). We will now formally introduce the optimization problem and recall that the corresponding decision problem is $\mathcal{NP}$-complete

- **Instance**: A finite set U, with $\mid U \mid = m$; a collection of sets $\mathcal{C} = \{S_1, \ldots, S_n\}$ such that $S_i \subseteq U \; \forall i = \{1, \ldots, n\}$. A weight function $w : U \to \Re^+$.
- **Solution**: A hitting set for $\mathcal{C}$, that is to say $H \subseteq U$ such that $H \cap S_i \neq \emptyset$, $\forall i = 1, \ldots, n$.
- **Optimal Solution**: a hitting set H such that $w(H) = \sum_{s \in H} w(s)$ is minimal.

The above definition is very general and, by simply putting $w(s) = 1$, $\forall s \in U$, we obtain the standard definition of the Minimum Hitting Set problem.

Theoretical results show (Feige, 1998) that optimal solutions cannot be approximated within $(1 - \epsilon) \ln m \; \forall \; \epsilon > 0$, unless $\mathcal{NP} \subset \mathcal{DTIME}(m^{\log \log m})$.

The Description of Our Algorithm. Our evolutionary approach and the resulting genetic algorithm, denoted by HEAT-V, is based on the idea of a *mutant virus*, which somehow acts as a non-purely random mutation operator. Each chromosome in the population is a binary string of fixed length (see below for details). The selection operator is *tournament selection* and the selected individuals mate with probability $p = 1$. Reproduction uses uniform crossover (however this does not involve the virus part as we will describe later). Elitism is used on three specific elements (not necessarily distinct) of the population: best fitness element; hitting set of smaller cardinality; and, hitting set of smaller weight.

Chromosomes contain some extra genes, specifically $2 + \lceil \log \mid U \mid \rceil$. These genes represent the genetic patrimony of the *virus*. As a consequence, the total length of a chromosome is $\mid U \mid +2 + \lceil \log \mid U \mid \rceil$. The extra $\lceil \log \mid U \mid \rceil$ bits uniquely identify one of the first $\mid U \mid$ *loci* of the chromosome. Viruses will hit an individual if the two extra control bits have both value 1.

We used three different fitness functions to test our algorithm. The best results were obtained using the function $f_3 : \mathcal{P} \to \mathcal{N} \setminus \{0\}$, that HEAT-V tries

E. Cantú-Paz et al. (Eds.): GECCO 2003, LNCS 2724, pp. 1570–1571, 2003.

to minimize, and which is defined as follows:

$$f_3(c) = w(c) + w(\mathcal{L}_{c,m})$$

where

$$\mathcal{L}_{c,m} = \{e : (\exists K \subseteq U) \text{ s.t. } K \cap c = \emptyset \wedge e \in K \wedge w(e) = \min\{w(e') : e' \in K\}\}.$$

Intuitively, f_3 is computed by adding to the weight of a chromosome, the minimum weight of elements of sets which c does not hit. Thus, f_3 acts as a strict upper-bound to the fitness function of any chromosome that could become a hitting set by including c.

Computational Results. We compared HEAT-V to a greedy algorithm which approximates the optimal solution to a factor of $O(\ln m)$. Basically, the procedure greedy chooses at every step the element that maximizes the ratio between the number of hit sets (among the remaining ones) and its weight. The hit sets are eliminated. Many tests were performed. For each test HEAT-V was tested three times. The population contained 200 individuals and each test ran for 500 generations. We also checked HEAT-V against *vertex cover*, which can be easily reduced to MHS. In particular, we used the regular graphs proposed by Papadimitriou and Steiglitz (PS-rg) (Papadimitriou and Steiglitz, 1982) built so that the classical greedy strategy fails. We ran HEAT-V with PS-rg's of degrees $k = 32, k = 66, k = 100$. HEAT-V always finds the optimal solution.

Conclusions and Future Work. We are now testing HEAT-V on a dynamic version of the WMHSP. More formally, we are dealing with cases in which with a given probability p_d a subset of the given family $\mathcal{C}$ disappears, and appears, instead, with probability p_a. Such a framework can be applied to many real scenarios, such as for instance, a computer network where each of the machine in the network offers specific services. Computers have a certain probability to go down, and for all of them the probability of going down is different, since it depends on factors such as the quality of the component, maintenance service, number of access, etc. To guarantee that the network is always able to provide a set of basic services, it is necessary to know which is the minimum number of computers that should be running in order to overcome some probabilistic failures. First experiments, conducted using Gaussian probability distributions, show that the population not only contains a good hitting set, but also good hitting sets in case some subsets disappear or appear.

References

1. Cutello, V., Mastriani, E., and Pappalardo, F. An evolutionary algorithm for the T-constrained variation of the Minimum Hitting Set Problem. In *Proceedings of 2002 IEEE Congress on Evolutionary Computation (CEC2002)* Vol. 1, pp. 366–371.
2. Feige, U. A threshold of $\log n$ for approximating set cover. *Journal of ACM 45* (1998), pp. 634–652.
3. Papadimitriou, C.H., and Steiglitz, K. *Combinatorial Optimization*. Prentice Hall (1982), p. 407.