Propositional SAT Algorithms and their Application to Solving HC Problems in Directed Graphs

Holger H. Hoos, Technical University of Darmstadt

– Summary of the Diploma thesis[1] –

The starting point of this work was the question, whether specific computational problems from molecular biology can be solved by application of powerful algorithms for propositional model finding to suitably encoded instances. This approach is based on the GSAT algorithm (Selman & al., 1992), a powerful, incomplete method for solving the NP-complete model finding problem in propositional logic (SAT) on the one hand side. On the other hand side, it is motivated by a recently published method of efficiently transforming Hamiltonion Circuit Problems (HCP) into SAT instances (Iwama & Miyzazaki, 1994).

In this work we show, how recognition of false hybridizations, a problem from the field of physical genome mapping that can be stated as NP-hard Traveling Salesman Problem (Karp, 1994) can be transformed into an optimization variant of SAT (MAX-SAT). We further explore, how these transformed problems can be solved by application of GSAT variants. The SAT transformation we use, is a variant of the HCP encoding proposed by Iwama & Miyazaki (see above). As it turns out, this approach is not practically applicable for realistic problem sizes; essentially this is caused by the transformation method.

Based on this result, in the main part of this work we explore properties of SAT encoded HC problems and the behaviour of GSAT-like algorithms on these in detail.

Accordingly, the following **questions** are discussed:

- Can GSAT be practically used to solve Iwama-Miyazaki-encoded Hamiltonion Circuit Problems in directed graphs (HCP)?
- Can its performance be improved by problem specific modifications of the basic algorithm?
- Which GSAT variants should be used to obtain the best performance?
- How can these algorithms' dependency of the performance on different parameter settings be characterized?
- Can optimal parameter settings be identified?
- How should the algorithms evaluated here be compared?
- Can GSAT be regarded as an adequate method for solving HCP??

[1] The Diploma thesis this summary is based on was carried out at the Comuter Science Department of the Technical University of Darmstadt under the supervision of Prof. Wolfgang Bibel und Antje Strohmaier (Intellectics Group).

From these questions, the following **main issues** of the thesis were derived:

- Characterization of structural properties of CNF-formulae which are obtained by the variable-efficient SAT-encoding of HCP instances by Iwama & Miyazaki;
- Analysis of the influence of these properties on the performance of GSAT-like algorithms;
- Problem-specific adaptions of GSAT to SAT-encoded HCP instances (GHC,GSHC);
- Empirical analysis of the runtime behaviour of GSAT, GHC, and GSHC on classes of HCP instances with different complexity (HC_1,HC_2,$HC_{n/2}$);
- Empirical evaluation of GSAT-variants (Adaptive Clause Weighting, Random Walk);
- Empirical evaluation of RBR-extension of the SAT encoded HCP instances (preconditioning by specific resolution mechanism);
- Empirical analysis of the influence of different parameter settings (`max-steps`, `walk-prob`) of GSAT, GHC, and GSHC on the performance of these algorithms.

The main results of the empirical part of this work are given as diagrams (see Appendix). We used three classes of HCP instances (HC_1, HC_2 und $HC_{n/2}$) and three algorithms for solving these (GSAT, GHC, and GSHC). Here, HC_k-n denotes the set of HCP instances with a graph consisting of n vertices, each of which has out-degree k. Furtheron, `max-steps` (ms) is the maximal number of steps per try for the given algorithms, `walk-prob` (wp) the probability for ding a step according to the Random Walk strategy, `succ-rate` (sr) the success rate of an (incomplete!) algorithm, and `steps/sln` (ss) the solutions costs (in steps/found solution).

In the following, we sum up the **main results** (references relate to the diagrams in the appendix):

- Iwama-Miyazaki encoded HCP instances show a structure that makes them particularly hard for GSAT-like algorithms (because of variable dependencies, the "score-by-need" method which is important for the performance of these algorithms is not effectively applicable here).
- By using Implicit Clause Encoding (ICE — motivated by a concept which was used by Selman & al. in a different context) the cubic space complexity for GSAT caused by the SAT-encoding method can be avoided, which is important for the practical application of this approach.
- The basic GSAT-algorithm was adapted to the specific structure of SAT-encoded HCP instances in two steps; thus two new algorithms GHC and GSHC were obtained. GHC is based on the idea of extending the neighbourhood relation which can be also applied to other, more general classes SAT instances. GSHC is a very HCP-specific method.
- Comparing the performance of these algorithms on identical test-sets of problems, GSHC is found to be better than GHC which in turn is significantly better than GSAT (cf. Fig.1-1a,1-2a,2-2a).
- A strong, qualitative influence of the Random Walk strategy on the performance of all three algorithms could be observed; the success-rate of each algorithm depends on the probability `walk-prob` for doing walk-steps (cf. Fig.1-1a,1-2a).

  For fixed `max-steps` values, the optimal setting for `walk-prob` is 0.5 for GSAT and GHC; for GSHC it is 0.2 on HC_1 problems, 0.4 for HC_2 (cf. Fig.1-6a,2-6a).

- Using Adaptive Clause Weighting does not have any significant influence on the performance of GSAT on SAT-encoded HCP instances.
- Applying Restricted Bound Resolution (RBR) extension to the problem instances as a preprocessing step does not affect GSAT's performance on these.
- As expected, HC_1 instances (solvable in linear time) are much easier than HC_2 instances (NP-hard). (This is a kind of problem specific adequateness.)

  Furthermore, $HC_{n/2}$ instances are much easier than HC_1 problems because they have a large number of solutions (cf. Fig.1-2a,2-2a).
- On HC_2 problems, for optimal settings of `walk-prob` we find an identical relation between `max-steps` and success-rate (sr/ms characteristic). This holds for all three algorithms independent of the problem size (cf. Fig.2-2a,2-2c,2-2e).

  Thus, GSAT,GHC, and GSHC are approximately complete in the sense that every solvable HCP instance can be solved with arbitrarily high probability (empirical result). This property can only be achieved by usings the Random Walk strategy (cf. Fig.1-1a,1-2a).

  The sr/ms characteristic can be approximated by a simple function: $sr(ms) = 1 - 2^{ms/c}$. Hence success-rates for arbitrary `max-steps` settings can be predicted; also for any given success-rate the `max-steps` value that is needed to obtain it can be derived (cf. Fig.2-3).

  The nature of the sr/ms characteristic implies that over a broad range doubling the `max-steps` value and doubling the number of tries yield indentical results. We therefore conclude that within this range the starting point for local search has almost no influence on the success-rate. (Consequently, in this situation one cannot expect to achive improvements by optimizing the initialization phase of local search.) Furtheron, this result implies that these algorithms can be optimally parallelized by simultanously running independent tries.
- For optimal `walk-prob` settings and arbitrarily chosen, but fixed success-rate, our experimental data indicates an exponential dependency of `max-steps` from the problem size (cf. Fig.2-4). Based on this result, we can predict the exact position of the sr/ms characteristic for arbitrary problem sizes. The predictions thus obtained could be confirmed for several sample instances.

  As a consequence, using these specific sr/ms characteristics, the relation between success-rate and problem size can be approximated for fixed `max-steps` values (cf. Fig.2-5).
- Based on these results, an upper bound for the run-time of the algorithms can be obtained for given problem size and success-rate; from this we can estimate the maximal size of practically solvable problem instances.

From these results several **open questions** arise some of which are addressed in the following:

- Is there a similar relation between `max-steps` and solution costs (steps per successful try) than the one we found between `max-steps` and success-rate (the sr/ms characteristic? (In this case one could obtain faithful run-time predictions for the algorithms.)
- Is it possible to characterize functional dependencies between the probability for Random Walk steps and the success-rate and/or solution costs? (Then we could predict the behaviour of the algorithm for the entire parameter space.)
- How can the observed relations between performance and parameter settings of the given algorithms be theoretically explained and which implications arise from such an explanation?
- Can the basic methodology and results of this investigation be applied to other problem classes (i.e. different classes of CNF formulae)? (By pursuing this question one could expect to gain insight into general properties of specific encoding methods and SAT algorithms.)
- How do complete SAT algorithms (like the latest, efficient variants of the Davis-Putnum procedures) perform on the class of SAT encoded HCP instances? (Until now it is largely unknown, for which problem classes local search methods like GSAT show similar or improved performance compared to "classic" backtracking algorithms.)
- Given the results of this work (esp. with respect to the SAT transformation method by Iwama & Miyazaki), what results can be achieved concerning the original problem from molecular biology and other "real world" optimization problems by following the general approach investigated here?