The Limits of Model-based Discovery?

Toby Walsh

University of York, York, England. tw@cs.york.ac.uk

Abstract. What are the limitations of model-based discovery? This paper identifies two limitations: limits of scale (models may have to be prohibitively large or numerous) and limits of sampling (individual models may mislead and generalization from them be difficult). Examples are used to illustrate both limits.

Introduction

Colton's HR program explores mathematical theories, inventing new concepts and making conjectures about these concepts [CBW99]. Otter is then called to prove these conjectures; if this fails, MACE is called to find a counter-example. For example, given some examples of groups of small order, HR invents the concept of Abelian group. Then, as all the examples of groups it has are Abelian, HR conjectures that all groups are Abelian. MACE is able to find a simple couter-example to this conjecture. This counter-example then enriches future concept formation as concepts and conjectures concerning non-Abelian groups can now be made.

HR has been very successful, discovering a number of integer sequences which are interesting and novel enough to have been accepted into the Encyclopedia of Integer Sequences [CBW00]. Whilst there are several reasons for this success, one of the most important is that HR is highly model driven. New concepts are made from old ones by forward chaining with a small number of production rules. To control the combinatorial explosion that results, concepts are assessed for their interestingness using the models at hand. For example, HR tests which examples of groups at hand have the new property. Concepts which are too specialized or too general are rated less interesting than concepts which partition the models at hand, especially when the partition is novel or close to some "gold standard" (like the classification of finite groups up to isomorphism). Similarly, inductive logic programming (ILP) is highly driven by models (the examples and counter-examples which we wish the program to cover). But what are the limits of such model-based discovery? In this paper, I present some examples which identify two possible limits.

Limits of scale

The size and number of models that must be considered to discover a particular concept might be prohibitively large. A recent example in number theory

illustrates this well. A year before his death in 1920, Ramanujan[1] came upon a remarkable pattern in the number of partitions of the first 200 integers. He conjectured and proved that $p(n)$, the number of partitions of the integer n obey the following congruences:

$$p(5n+4) = 0 \bmod 5$$
$$p(7n+5) = 0 \bmod 7$$
$$p(11n+6) = 0 \bmod 11$$

Similar congruences occur where the interval between integers is a power of 5, 7, or 11 or a product of these powers.

Many decades of largely fruitless searching yielded just one or two more apparently isolated congruences. It was threfore generally believed that few other congruences existed. Recently, however, Ono has proved that congruences exist for all larger primes [Ono00]. His research involved almost no computation and relied heavily on the theory of modular forms. Subsequent work by a student of Ono's found a simple criterion for identifying the start of a progression. This gave an algorithm with which more than 70,000 new congruences have been found including:

$$p(11,864,749n+56,062) = 0 \bmod 13$$
$$p(14,375n+3,474) = 0 \bmod 23$$

Given the size of the coefficients in these congruences, it is unsurprising that mathematicians had previously failed to find them. However, it is also unlikely that a very fast computer program could have found them by looking for patterns in a suitable database. The program would have had to cast its net very wide, which is both expensive and likely to throw up false leads. Although congruences like these are "plentiful", they are typical very very large. Another example of a concept that would defeat a model-based approach concerns simple groups, groups with no non-trivial normal subgroup. The smallest non-cyclic normal subgroup is of order 60. A model-based program like HR might therefore conjecture that all simple groups are cyclic based on the smaller order groups to hand. However, it would be beyond the capabilities of programs like MACE to find a couter-example to this conjecture[2].

These examples suggest that model-based discovery will face limits of scale; models may have to be prohibitively large or numerous for discoveries to be made. Techniques based on a deeper understanding of the structure of the theory may therefore have an important role to play.

Limits of sampling

Even for finite systems, model-based discovery may not be able to sample enough of the space or may be mislead by individual examples. Consider the discovery of game playing strategies for games with incomplete information such as

[1] HR is named in honour of Hardy and Ramanujan.

[2] This example was suggested by Ursula Martin at the Calculemus workshop in 1998

bridge. Programs like Ginsberg's GIB [Gin99] use Monte-Carlo sampling to guess "likely" states of the game, and then standard complete information methods like Min-Max or Alpha-Beta are used to find a good move. The hope is that by examining a representative sample of possible worlds, a move that works well in a large percentage can be identified. However, Frank and Basin have shown that, even for very simple game trees, the chance of finding an optimal strategy with Monte-Carlo sampling rapidly approaches zero as the size of the tree increases [FBM98,FB98]. Two problems arise. The first problem is "strategy fusion", the problem of combining strategies from different worlds to produce optimal strategy across all worlds. Unfortunately, the most common best strategy for a single world is not necessarily the best across all worlds. This may be seen as a problem of generalization. The second problem is "non-locality", the problem that an opponent with partial knowledge of the game can direct play to more favourable parts of game tree, avoiding those parts of game tree which are particularly bad.

Thls example suggests that model-based discovery will face limits of sampling; individual models may mislead and generalization from them be difficult. As with limits of scale, a deeper understanding of the structure of the theory may have an important role to play in such cases.

Conclusions

Despite the successes of model-based discovery, I believe that there may be a number of limitations on making discoveries guided purely by models. In this paper, I identify two such limitations: limits of scale (models may have to be prohibitively large or numerous) and limits of sampling (individual models may mislead and generalization from them be difficult).

Acknowledgements

The author is supported by an EPSRC advanced research fellowship.

References

[CBW99] S. Colton, A. Bundy, and T. Walsh. Automatic identification of mathematical concepts. In *Proceedings of 16th IJCAI*. International Joint Conference on Artificial Intelligence, 1999.

[CBW00] S. Colton, A. Bundy, and T. Walsh. Automatic invention of integer sequences. In *Proceedings of the 17th National Conference on AI*. American Association for Artificial Intelligence, 2000.

[FB98] I. Frank and D. Basin. Search in games with incomplete information: A case study using bridge card play. *Artificial Intelligence*, 100(1-2):87–123, 1998.

[FBM98] I. Frank, D. Basin, and H. Matsubara. Finding optimal strategies for imperfect information games. In *Proceedings of the 15th National Conference on AI*, pages 500–507. American Association for Artificial Intelligence, 1998.

[Gin99] M. Ginsberg. Gib: Steps towards an expert-level bridge-playing program. In *Proceedings of the 16th IJCAI*. International Joint Conference on Artificial Intelligence, 1999.

[Ono00] K. Ono. Distribution of the partition function modulo m. *Annals of Mathematics*, 151:293–307, 2000.