Analysis and Improvement of Genetic Algorithms using Concepts from Information Theory

John Milton

THESIS

Submitted to the Faculty of Engineering and IT
University of Technology Sydney (UTS)
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Computing Sciences

2009

CERTIFICATE OF AUTHORSHIP/ORIGINALITY

I certify that the work in this thesis has not previously been submitted for a degree nor has it been submitted as part of requirements for a degree except as fully acknowledged within the text.

I also certify that the Thesis has been written by me. Any help that I have received in my research work and the preparation of the thesis itself has been acknowledged. In addition, I certify that all information sources and literature used are indicated in the thesis.

Signature of Candidate

ACKNOWLEDGEMENT

The patience of my wife and son have been greatly appreciated over the many years of work required to complete this Thesis part-time.

I would also like to thank Dr Paul Kennedy for his advice, patience and tireless editorial assistance with both this Thesis and the papers which preceded some of the chapters. Paul's professionalism and guidance have been of enormous benefit to me, both in seeing this thing through and in the quality of the research undertaken.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABSTRACT

Evolutionary algorithms are based on the principles of biological evolution (Bremermann et al., 1966; Fraser, 1957; Box, 1957). Genetic algorithms are a class of evolutionary algorithm applicable to optimisation of a wide range of problems because they do not assume that the problem to be optimised is differentiable or convex. Potential solutions to a problem are encoded by allele sequences (genes) on an artificial genome in a manner analogous to biological DNA. Populations of these artificial genomes are then tested and bred together, combining artificial genetic material by the operation of crossover and mutation of genes, so that encoded solutions which more completely optimise the problem flourish and weaker solutions die out.

Genetic algorithms are applied to a very broad range of problems in a variety of industries including financial modeling, manufacturing, data mining, engineering, design and science. Some examples are:

- Traveling Salesman Problems such as vehicle routing,
- Scheduling Problems such as Multiprocessor scheduling, and
- Packing problems such as Shipping Container Operations.

However, relative to the total volume of papers on genetic algorithms, few have focused on the theoretical foundations and identification of techniques to build effective genetic algorithms. Recent research has tended to focus on industry applications, rather than design techniques or parameter setting for genetic

algorithms. There are of course exceptions to these observations. Nevertheless, the exceptions generally focus on a particular parameter or operator in relative isolation and do not attempt to find a foundation, approach or model which underpins them all.

The objective of this Thesis is to establish theoretically sound methods for estimating appropriate parameter settings and structurally appropriate operators for genetic algorithms. The Thesis observes a link between some fundamental ideas in information theory and the relative frequency of alleles in a population. This observation leads to a systematic approach to determining optimum values for genetic algorithm parameters and the use of generational operators such as mutation, selection, crossover and termination criteria. The practical significance of the Thesis is that the outcomes form theoretically justified guidelines for researchers and practitioners.

The Thesis establishes a model for the analysis of genetic algorithm behaviour by applying fundamental concepts from information theory. The use of information theory grounds the model and contributions to a well established mathematical framework making them reliable and reproducible. The model and techniques contribute to the field of genetic algorithms by providing a clear and practical basis for algorithm design and tuning.

Two ideas are central to the approach taken. Firstly, that evolutionary processes encode information into a population by altering the relative frequency of alleles. Secondly, that the key difference between a genetic algorithm and other algorithms is the generational operators, selection and crossover. Hence the model maximises a population's information as represented by the relative frequency of solution alleles in the population, encourages the accumulation of these alleles and maximises the number of generations able to be processed.

Information theory is applied to characterise the information sources used for mutation as well as to define selection thresholds in ranked populations. The importance of crossover in distributing alleles throughout a population and in promoting the accumulation of information in populations is analysed, while the Shannon–McMillan theorem is applied to identify practical termination criteria.

The concept of *ideal alleles* as being those symbols in the appropriate loci, which form an optimal solution and the associated *solution density* of the population is central to this analysis. The term *solution density* is introduced to refer to the relative frequency of ideal alleles in the population at a particular generation. Solution density so defined represents a measure of a population's fitness.

By analysing the key genetic operators in terms of their effect on solution density, this Thesis identifies ten contributions.

- A model for the analysis of genetic algorithm behaviour inspired by information theory.
- A *static selection* threshold in ranked populations.
- A *dynamic selection* threshold in ranked populations.
- A maximum limit on the number of loci participating in epistasis is identified whereby more epistatic loci degrade directed search.
- A practical limit to the amount of useful crossover is identified as *sufficient*.
- An optimal crossover section length is found.
- A cumulative scoring method for identifying solution density.
- An *entropy profile* of ranked lists is described.
- A practical termination criteria of *most probable individuals* based on the Shannon–McMillan theorem is provided.

- An alternative genome representation which incorporates job–shop schedule problem knowledge in the genome rather than the algorithm's generational operators is developed.

Each of these contributions is validated by simulations, benchmark problems and application to a real–world problem.