Using Expert Knowledge in Initialization for Genome-wide Analysis of Epistasis Using Genetic Programming

Casey S. Greene
Dartmouth College
Lebanon, NH 03756 USA

Bill C. White
Dartmouth College
Lebanon, NH 03756 USA

Jason H. Moore
Dartmouth College
Lebanon, NH 03756 USA
Jason.H.Moore@dartmouth.edu

ABSTRACT

In human genetics it is now possible to measure large numbers of DNA sequence variations across the human genome. Given current knowledge about biological networks and disease processes it seems likely that disease risk can best be modeled by interactions between biological components, which may be examined as interacting DNA sequence variations. The machine learning challenge is to effectively explore interactions in these datasets to identify combinations of variations which are predictive of common human diseases. Genetic programming is a promising approach to this problem. The goal of this study is to examine the role that an expert knowledge aware initializer can play in the framework of genetic programming. We show that this expert knowledge aware initializer outperforms both a random initializer and an enumerative initializer.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: biology and genetics

General Terms

Algorithms, Performance

Keywords

Genetic Analysis, Genetic Programming, Expert Knowledge, Initialization

1. INTRODUCTION

In human genetics it is now possible to measure more than one million DNA sequence variations from across the human genome. An important goal in human genetics is the determination of which of the variations are useful for predicting individual risk for common diseases. Because of the complexity of biological networks, epistasis, which describes nonlinear interactions between genes, is likely to be ubiquitous. Combining the difficulty of modeling nonlinear attribute interactions with the challenge of attribute selection yields for this domain what Goldberg [1] calls a needle-in-a-haystack problem. There may be a particular combination of attributes that together with the right nonlinear function are a predictor of disease susceptibility. Considered individually they may not look any different than thousands of other noisy attributes not involved in the disease process.

Genetic programming (GP) is an automated computational discovery tool that is inspired by Darwinian evolution and natural selection. This is accomplished by first generating computer programs that are composed of the building blocks needed to solve or approximate a solution to a problem. Each generated program is evaluated, and the good programs are selected, recombined, and mutated to form new computer programs. Genetic programming and its many variations have been applied successfully to a wide range of different problems. Work here examines whether or not it is possible to use expert knowledge in initialization to develop a GP strategy which performs better than one with a standard initialization operator.

Work on initialization in GP has largely centered on the problem of generating diverse and valid tree structures without overwhelming computational complexity. O'Neill and Ryan [4] discuss the importance of initialization and the impact of diversity on final solutions. Here we apply their principles of sensible initialization through an exhaustive initializer focused on diversity and an expert knowledge based initializer focused on exploiting knowledge about the problem to population initialization.

2. AN EXPERT KNOWLEDGE AWARE INITIALIZATION OPERATOR

The goal of this study was to examine whether expert knowledge could be used to ensure good building blocks are introduced into the population through initialization. We compared three initializers in this study. All of these initializers create a tree with the MDR function as the root node and two attributes (SNPs) as the leaves. The first initializer is a random initializer. The attributes chosen as leaves are selected randomly from the list of available attributes. The second initializer is an exhaustive initializer. All available attributes are stored in a vector. The vector is shuffled and the attributes chosen as leaves are selected successively from the shuffled vector. When the end is reached the vector is reshuffled and the process begins at the beginning again. The third initializer is an expert knowledge aware probabilistic initializer. Attributes are selected as leaves via a roulette wheel approach using TuRF scores to prepare the roulette wheel. The same attribute is not allowed to be used twice within the same tree, but it may be used any number of times within the generated population.

GECCO'08, July 12–16, 2008, Atlanta, Georgia, USA.
ACM 978-1-60558-130-9/08/07.

2.1 Parameter Settings

For this study, we use a population size of 5000 and run the GP for 10 generations. We use a crossover probability of 0.9 and no mutation. Since each tree has exactly two attributes, an initial population size of 5000 trees will include 10,000 total attributes. The initial population was generated using one of the three initializers. Runs were performed both with and without the use of expert knowledge in the fitness function.

3. MULTIFACTOR DIMENSIONALITY REDUCTION (MDR) FOR ATTRIBUTE CONSTRUCTION

Multifactor dimensionality reduction (MDR) was developed as a nonparametric and genetic model-free data mining strategy for identifying combination of SNPs that are predictive of a discrete clinical endpoint. The MDR method has been successfully applied to detecting gene-gene interactions for a variety of common human diseases.

4. EXPERT KNOWLEDGE FROM TUNED RELIEFF (TURF)

Our goal is to provide an external measure of attribute quality that can be used as expert knowledge for population initialization by the GP. Kira and Rendell [2] developed an algorithm called Relief that is capable of detecting attribute dependencies. Kononenko improved upon Relief by choosing n nearest neighbors instead of just one. This new ReliefF algorithm has been shown to be more robust to noisy attributes and missing data [5] and is widely used in data mining applications. We have developed a modified ReliefF algorithm for the domain of human genetics called Tuned ReliefF (TuRF).

5. DATA SIMULATION AND ANALYSIS

The goal of the simulation study is to generate artificial datasets with high concept difficulty to evaluate the power of GP in the domain of human genetics. We develop 30 different penetrance functions (i.e. genetic models) that define a probabilistic relationship between genotype and phenotype where susceptibility to disease is dependent on genotypes from two SNPs in the absence of any independent effects. The penetrance functions include heritabilities of 0.025, 0.05, 0.1, 0.2, 0.3, or 0.4. Each functional SNP has two alleles with frequencies of 0.4 and 0.6.

For each set of 100 datasets and for each set of parameters we count the number of times the correct two functional attributes are selected as the best model by the GP. This count, expressed as a percentage, is an estimate of the power of the method. We compare the significance of power estimates between the methods (e.g. exhaustive initializer vs expert knowledge initializer) by performing a chi-square test of independence. Results are considered statistically significant when the p-value for the chi-square test statistic was ≤ 0.05.

6. EXPERIMENTAL RESULTS

The exhaustive initializer did not perform differently than the random initializer in most cases ($p > 0.05$). In contrast across all heritabilities for a major allele frequency of 0.6 the expert knowledge aware initializer was significantly different than both the exhaustive and random initializers ($p < 0.05$) when expert knowledge was also used in the fitness function. When expert knowledge was not used in the fitness function the expert knowledge initializer significantly differed from the exhaustive initializer and random initializer across all tested heritabilities when the major allele frequency was 0.6 and across heritabilities ($p < 0.05$). This is clear evidence that the expert knowledge initializer provides the rest of the GP operators with a population containing many good building blocks.

7. DISCUSSION AND CONCLUSION

Firstly, we have shown that expert knowledge can provide building blocks necessary to find the genetic needle in the genome-wide haystack. Secondly, expert knowledge aware initialization performs better than both random initialization and exhaustive initialization. The initialization method makes a significant difference in the outcome confirming O'Neill and Ryan's suggestion [4] that the initial population has a large impact on the outcome. In addition the results using a simple fitness function which integrates TuRF scores show that the expert knowledge initializer also greatly increases the success when other expert knowledge features are added to the GP. Combining this initializer with other knowledge guided strategies in selection, mutation, and recombination may provide additional benefits.

Moore et al. have recently shown that Symbolic Discriminant Analysis (SDA), which uses a GP approach to generate models, was able to successfully model predictors of atrial fibrillation in a well characterized dataset which included a two-way epistatic interaction [3]. Integrating expert knowledge into the SDA approach should increase the efficiency of the search, assisting SDA in finding higher order interactions and allowing SDA to be applied to larger datasets. One attractive feature of the probabilistic initializer is that it is easily integrated into already existing approaches. This study brings us one step closer to routine use of GP strategies for the genetic analysis of common human diseases.

8. ACKNOWLEDGMENTS

This work was supported by NIH grants LM009012 and AI59694.

9. REFERENCES

[1] D. E. Goldberg. *The Design of Innovation: Lessons from and for Competent Genetic Algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 2002.

[2] K. Kira and L. A. Rendell. A practical approach to feature selection. *In: Machine Learning: Proceedings of the AAAI'92*, 1992.

[3] J. H. Moore, N. Barney, C. T. Tsai, F. T. Chiang, J. Gui, and B. C. White. Symbolic modeling of epistasis. *Hum Hered*, 63(2):120–133, Feb 2007.

[4] M. O'Neill and C. Ryan. *Grammatical Evolution: Evolutionary Automatic Programming in an Arbitrary Language*. Kluwer Academic Publishers, Norwell, MA, USA, 2003.

[5] M. Robnik-Sikonja and I. Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Mach. Learn.*, 53(1-2):23–69, 2003.