A Comparison of GE Optimized Neural Networks and Decision Trees

Kristopher Hoover*
NC State Unversity Institute for Advanced Analytics
Raleigh, North Carolina 27606
kmhoover@ncsu.edu

Rachel Marceau*
NC State University
Raleigh, North Carolina 27695
remarcea@ncsu.edu

Tyndall Harris
NC State University
Raleigh, North Carolina 27695
tpharris@ncsu.edu

David Reif
NC State University
Raleigh, North Carolina 27695
dmreif@ncsu.edu

Alison Motsinger-Reif
NC State University
Raleigh, North Carolina 27695
aamotsin@ncsu.edu

ABSTRACT
Grammatical evolution neural networks (GENN) is a commonly utilized method at identifying difficult to detect gene-gene and gene-environment interactions. It has been shown to be an effective tool in the prediction of common diseases using single nucleotide polymorphisms (SNPs). However, GENN lacks interpretability because it is a black box model. Therefore, grammatical evolution of decision trees (GEDT) is being considered as an alternative, as decision trees are easily interpretable for clinicians. Previously, the most effective parameters for GEDT and GENN were found using parameter sweeps. Since GEDT is much more intuitive and easy to understand, it becomes important to compare its predictive power to that of GENN. We show that it is not as effective as GENN at detecting disease causing polymorphisms especially in more difficult to detect models, but this power trade off may be worth it for interpretability.

Categories and Subject Descriptors
I.6.4 [**Simulation and Modeling**]: Model Validation and Analysis

General Terms
Algorithms, Performance, Verification.

Keywords
Grammatical evolution, neural networks, decision trees, genetic algorithms

1. INTRODUCTION
With advancing technology, geneticists are able to collect and store more and more data, giving them further insight into the complex relationships in which genes interact to cause diseases. Previously all diseases were believed to be "common" – that is, easily explained by few genes without many interactions. In these diseases, traditional statistical methods such as logistic regression and the chi squared test were used to determine which genes were responsible for disease status in an individual. However, today, as we are able to sequence more individuals and more of the genome, it has become apparent the majority of diseases are in fact complex and epistatic in nature, where many genes interact with each other and with the environment to cause diseases [1]. In these cases, traditional statistical methods have failed to predict disease status.

A new method that is becoming more popular in the field of genetics is Evolutionary Computation (EC). The method of EC mimics natural selection, allowing a population of statistical models to "breed" and "mutate" where only the best models survive [2].

One of the most common EC algorithms used in genetics is Grammatical Evolution (GE) which utilizes a set of rules known as grammars to properly evolve the population of models. GE has been explored on many different families of models, including petri-nets, neural networks, and decision trees. Each has been studied separately in the past, and the optimal parameter settings for each type of model was found [3, 9]. Here we focus on comparing Grammatical Evolution of Neural Networks (GENN) and Grammatical Evolution of Decision Trees (GEDT), as they are the most promising GE models currently available [4].

2. METHODS

2.1 Grammatical Evolution
Grammatical Evolution is a type of genetic algorithm which mimics natural selection using a set of rules called a grammar. Grammatical evolution initializes with a population of individual models, each with their own starting fitness. This starting fitness is a direct comparison of how well the newly created models correlate to the true model [5]. Cross-validation is used to help determine the fitness of a model: models can be created and tested using 9/10 of the data, then their predictive error can be calculated using the additional 1/10 of the data in a 10-fold cross validation [6]. The population of models competes against each other, via a predetermined selection method (tournament, roulette, etc.), to determine which models the "fittest" and thus can pass on their genetic model material. For each generation the fittest are either passed onto the next generation as is, can mutate and be passed on, or can breed with another model ("cross-over"), combining to create a new model in the next generation, all with a determined probability. Fittest must be defined but is often considered those models with the smallest error. This selection process continues

GECCO'12 Companion, July 7–11, 2012, Philadelphia, PA, USA.

until a specified number of generations have passed, or until a certain fitness level is obtained [7].

2.2 Grammatical Evolution of Decision Trees (GEDT)

Alone, decision trees are hierarchical and do not allow for interaction between variables [8]. However, grammatical evolution can be applied to decision trees, using the above mentioned method of natural selection to optimize the decision trees. Decision trees are created with nodes displaying loci chosen to be in the model, each with three branches: one for each possible genotype at the loci. Each branch then shows whether the simulated individual has the disease of interest, i.e. is a case, or does not and is a control, or, in epistatic models, will branch to another loci.

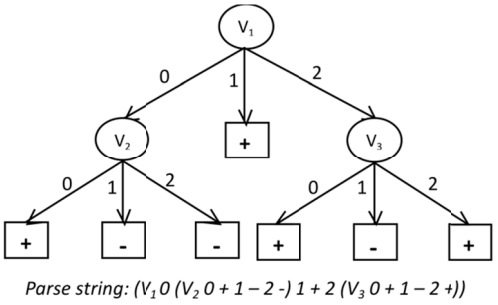

Figure 1. Decision Tree [adapted from 8]
SNPs are found at the nodes, which branch out depending on alleles at that SNP to give ultimate disease status depending on the interactions of these SNPs.

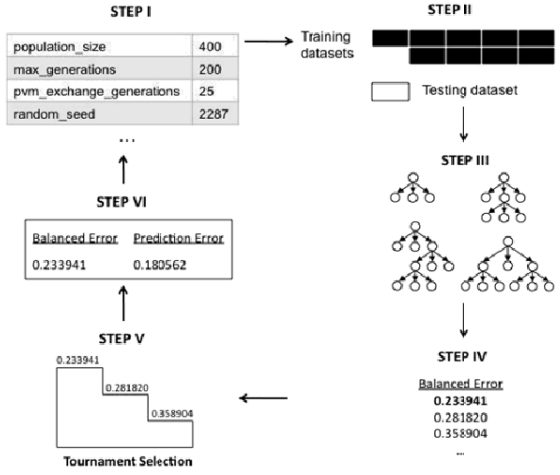

Figure 2. Grammatical Evolution of Decision Trees
After parameter initialization, GEDT takes 9/10 of the data and creates decision trees which are evaluated, and the ones with the least error are evaluated on the remaining 1/10 of the data. These models go on and mate to create the next generation of models [6].

2.3 Grammatical Evolution of Neural Networks (GENN)

Unlike decision trees, neural networks are black box models. That is, you are given the input and output, but cannot see what is happening between the two. In modeling genetic diseases, you have the input, the individuals' SNPs, and the output, the disease status, but can't determine which SNPs interact to create the disease or in what fashion they do.

Using grammatical evolution with neural networks has been shown to be a very effective combination at identifying gene-gene interactions. GENN has a very high statistical power, however it creates models which are not easily interpretable [9]. This makes it difficult for pharmacists and nurses to understand what the model actually means.

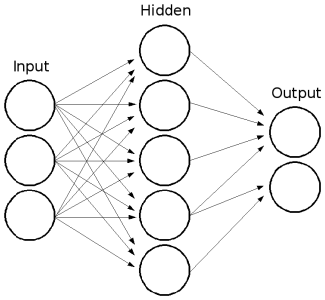

Figure 3. Neural Network
SNPs are input, and follow through a hidden model to give disease status.

2.4 Datasets

Multiple datasets which attempted to mimic real life genetic data were used. Complete description of their simulation can be found in [9]. Each dataset used had 500 cases and 500 controls where each individual had 100 SNPs. Data was generated such that two of these SNPs were disease-causing with significant interaction between the two, and the other 98 were noise loci. This epistatic relationship is much more difficult to find than the common disease case in which there is no interaction, so if the methods can perform well in the epistatic case, they are expected to perform well overall.

Different levels of heritability (5%, 10%, 15%, 20%, and 25%) and minor allele frequency (0.2, and 0.4) were modeled, and all combinations were considered to create 10 total models of interest. 100 datasets were used for each model, resulting in 1000 total datasets. Models with higher heritability should be easier to find, because more of the variability is due to genetics rather than the environment, which our data does not contain information on. The minor allele frequency is how often the rare allele is seen in the population.

2.5 Implementation

Parameters used were as suggested by a previous study intending to optimizing the efficiency of GEDT [3], and GENN [9]. For GEDT, we let the simulations run for 550 generations with an initial population size of 750 models, choosing parents using a tournament selection. We let the data crossover 80% of the time,

and mutate 5% of the time. For GENN, simulations were run for 400 generations with an initial population of 200 models. The data had a 90% crossover rate, and a 1% mutation rate, with tournament selection. Additional GA parameters used are as follows:
codon size = 8, GE wrapping count = 2, min chromosome size = 50, max chromosome size = 1000, and sensible initialization depth = 10 [9].

Most of the simulations were run on a server running Ubuntu Linux or on N.C. State University's High Performance Computing (HPC) Cluster, which uses single, dual, and quad core Xeon Processors with up to 3 GB of memory each [http://www.ncsu.edu/itd/hpc/Hardware/Hardware.php].

2.6 Power Calculations

To calculate the effectiveness of each method, cross validation was used. Cross validation is a method that creates a model using only 90% of the data and then tests the model to the other 10% of data. This process is repeated until all of the data had been tested. The purpose of cross validation is to limit over-fitting the model. Therefore, only loci with cross validations high enough to be considered statistically significant were included in the model. Power was defined as the number of times the method correctly identified the actual disease-causing loci as the top 2 loci in the cross-validation procedure out of each set of 100 datasets.

3. RESULTS

Table 1. Results from 100 SNP, 1000 individual Purely Epistasic Models

Model			Power (%)	
# Functional SNPs	Allele Frequencies	Heritability	GENN	GEDT
2	0.2/0.8	5%	66	55
2	0.2/0.8	10%	87	65
2	0.2/0.8	15%	99	85
2	0.2/0.8	20%	95	87
2	0.2/0.8	25%	88	88
2	0.4/0.6	5%	98	78
2	0.4/0.6	10%	99	84
2	0.4/0.6	15%	100	90
2	0.4/0.6	20%	94	91
2	0.4/0.6	25%	100	92

An ANOVA was done to look at the effect of major allele frequency (or equivalently minor allele frequency), (percent) heritability, and type (GENN or GEDT) on power. All were found to have a significant effect (with p-values of 0.0056, 0.0027, and 0.0056, respectively). It had been previously found allele frequency and heritability were important for both methods [3,9], however, we found that type did have an important effect on power at the alpha = 0.01 level after accounting for heritability and allele frequency. GENN was found to be significantly more powerful than GEDT at detecting epistasis. Fitting an interaction model suggested that all factors were still important as above (p = 0.0031, 0.0016, 0.0031, respectively), but all two-way interaction terms were not significant. It is interesting to note that the three-way interaction term was significant at the 0.1 level (p=0.0858). This is visible in the graph below.

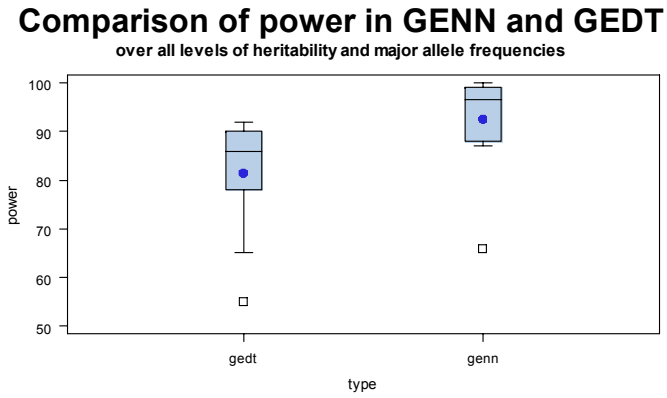

Figure 4. Boxplots comparing the power of GENN and GEDT summarized over all parameter settings considered

Here we see GENN is considerably more powerful than GEDT. It is estimated that using GEDT rather than GENN, on average, will produce an 11 point decrease in power (estimate=-11.1, std error= 3.475, p=0.0056).

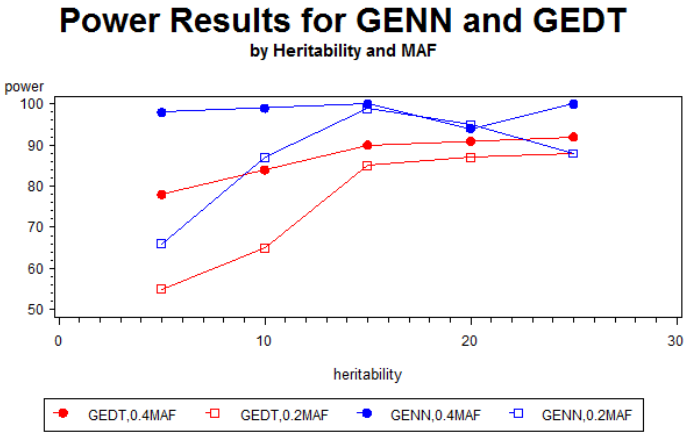

Figure 5. Plot showing the power of GENN and GEDT for each heritability and minor allele frequency combination

The difference in power appears to be smaller for larger levels of heritability and smaller MAF.

4. DISCUSSION

Ultimately, it seems as though GENN is in fact a better option, because of its significant increase in power. GEDT would likely be used only when interpretability is so important you are willing to sacrifice power, though this difference in power is less extreme for large levels of heritability and larger major allele frequencies. This is sensible as higher heritabilities are easier to detect for reasons listed above, and smaller minor allele frequencies are again more extreme and could be easier to detect. Thus, GEDT is a good alternative if you expect the true model to be easier to detect. In difficult cases, GENN should be used.

It is important to continue this test on different levels of heritability and minor allele frequency to see if this is a general pattern.

5. ACKNOWLEDGMENTS

The research is based upon work supported by the National Science Foundation under CSUMS grant #DMS-0703392 (PI: Sujit Ghosh). The authors would like to thank the other participants of the Computation for Undergraduates in Statistics Program for their helpful input into the project, including Sujit Ghosh, Jackie Dietz, Olivia Bagley, Allison Huber, Wesley Stewart, Rachael Beckner, Nicole Mack, and James Kniffen.

6. REFERENCES

[1] Moore, J.H., *The ubiquitous nature of epistasis in determining susceptibility to common human diseases.* Hum Hered, 2003. 56(1-3): p. 73-82.

[2] Motsinger, A.A., M.D. Ritchie, and D.M. Reif, *Novel methods for detecting epistasis in pharmacogenomics studies.* Pharmacogenomics, 2007. 8(9): p. 1229-41.

[3] Hoover, Kristopher, Rachel Marceau, Tyndall Harris, Nicholas Hardison, David Reif, and Alison Motsinger-Reif. "Optimization of Grammatical Evolution Decision Trees." *GECCO '11: Proceedings of the 13th Annual Conference Companion on Genetic and Evolutionary Computation*. Dublin, Ireland. New York: ACM, 2011. 35-36.

[4] O'Neill, M. and C. Ryan, *Grammatical Evolution: Evolutionary automatic programming in an arbitrary language.* 2003, Boston: Kluwer Academic Publishers

[5] Hastie, T.J., R.J. Tibshirani, and J.H. Friedman, *The elements of statistical learning*. Springer Series in Statistics. 2001, Basel: Springer Verlag.

[6] Miller, B.L.G., D.E., *Genetic Algorithms, Tournament Selection and the Effects of Noise.* Complex Systems, 1995. 9(3): p. 193-212.

[7] Alpaydin, E., *Introduction to Machine Learning*. 2004, Cambridge, MA: MIT Press.

[8] Motsinger-Reif, A.A., et al., *Grammatical evolution decision trees for detecting gene-gene interactions.* BioData Min, 2010. 3(1): p. 8.

[9] Motsinger-Reif, A. A., Dudek, S. M., Hahn, L. W. and Ritchie, M. D. (2008), *Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology*. Genet. Epidemiol., 32: 325–340. doi: 10.1002/gepi.203.