Grammatical Evolution Support Vector Machines for Predicting Human Genetic Disease Association

Skylar Marvel
North Carolina State University
Bioinformatics Research Center
Raleigh, NC 27695
swmarvel@ncsu.edu

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

ABSTRACT

Identifying genes that predict common, complex human diseases is a major goal of human genetics. This is made difficult by the effect of epistatic interactions and the need to analyze datasets with high-dimensional feature spaces. Many classification methods have been applied to this problem, one of the more recent being Support Vector Machines (SVM). Selection of which features to include in the SVM model and what parameters or kernels to use can often be a difficult task. This work uses Grammatical Evolution (GE) as a way to choose features and parameters. Initial results look promising and encourage further development and testing of this new approach.

Categories and Subject Descriptors

I.2.m [**Artificial Intelligence**]: Miscellaneous—*Genetic-Based Machine Learning and Learning Classifier Systems*

General Terms

Algorithms

Keywords

Support vector machine, grammatical evolution, Single Nucleotide Polymorphism (SNP), epistasis

1. INTRODUCTION

The ability to identify genes that predict common, complex human diseases is an intense area of research. Such diseases are often caused by the combination of many genetic and environmental factors, each contributing a small effect [8]. Identification of genetic factors is made difficult by the interactions between different genes, referred to as epistasis [3]. Traditional parametric statistical methods used to characterize gene-gene or gene-environment interactions fail when applied to large datasets [4], which has stimulated the development of novel computational approaches that are able to extract information from data obtained during this 'omics' era.

One popular approach for detecting disease association involves the use of machine-learning classification methods [1, 2, 9, 14]. A few of the most common methods are Artificial Neural Networks (ANNs), Decision Trees (DTs) and Support Vector Machines (SVMs), the later of which has been steadily gaining popularity. Due to the enormous size of the datasets that are being analyzed, feature selection is an extremely important aspect of these classification methods [11]. In addition, properties innate to the classification technique also influence performance, e.g. the architecture of an artificial neural network or the kernel parameter(s) of a support vector machine.

To address these issues, many techniques are being developed that combine machine-learning classification methods with algorithms that select features and classifier architecture [2, 7, 9, 12]. Genetic programming algorithms are often used for this purpose [2, 7, 12], however, application of Grammatical Evolution (GE) has been shown to outperform the genetic programming counterpart for ANNs [9]. Motivated by this result and the increasing use of SMVs, this work begins the process of combining GE and SVMs for the purpose of predicting human genetic disease associations.

2. METHODS

2.1 Support Vector Machines

SVMs are non-probabilistic binary classifiers that can be used to construct a hyperplane to separate data into one of two classes [13]. Consider a set of n data points, each consisting of p features, $\mathbf{x} \in \mathbb{R}^p$, and a class label, $y \in [-1, 1]$, i.e. $(\mathbf{x}_i, y_i)$ for $i = 1, \ldots, n$. A hyperplane can be defined by a normal vector, $\mathbf{w}$, and offset, b. In addition, slack variables, ξ_i, can be introduced to represent the degree of misclassification when data points are not linearly separable. The objective function of the SVM is then

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{n}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0, \end{aligned} \tag{1}$$

where C is a linear misclassification penalty and ϕ is a non-linear transformation function that projects $\mathbf{x} \in \mathbb{R}^p$ into a higher-dimensional feature space. Using the relationship

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

$\mathbf{w} = \sum_{i=1}^{n} \alpha_i y_i \mathbf{x}_i$, this problem can be solved in dual form

$$\begin{aligned} \max_{\boldsymbol{\alpha}} \quad & \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C, \\ & \sum_{i=1}^{n} \alpha_i y_i = 0, \end{aligned} \tag{2}$$

where $k(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i) \cdot \phi(\mathbf{x}_j)$ is a nonlinear kernel function. One of the most popular kernel functions, and the one used in this paper, is the radial basis function (RBF) $k(\mathbf{x}_i, \mathbf{x}_j) = e^{-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2}$, for $\gamma > 0$ [2].

2.2 Grammatical Evolution

Grammatical Evolution (GE) is a technique similar to genetic programming that is capable of exploring vast search spaces involving potential features and classification architectures. Details of GE can be found in O'Neill and Ryan [10]. GE uses a Backus-Naur Form grammar to convert a binary string into a computer program in any language by following the central dogma of biology where a binary string (genotype) is converted to an integer string (mRNA) and then mapped to a program (phenotype). The use of a grammar allows for unconstrained searching via evolutionary operations to occur at the genetic level, such as point mutations and crossover events, while maintaining a valid phenotypic program. GE has been used successfully in the creation of ANNs and DTs for identifying disease susceptibility genes [9]. The use of genetic algorithms to optimize the feature selection and architecture of both ANNs and SVMs has been investigated [2, 7, 12], however, to our knowledge GE has not been applied to SVMs.

2.3 Grammatical Evolution Support Vector Machine (GESVM)

In the current study we have developed a method that uses GE to select features and parameter values for a SVM classifier that can be applied to case/control data for human diseases. We have initially selected the RBF kernel, which requires a value for the kernel parameter γ in addition to the misclassification penalty C. We designed the GESVM for use with datasets where the features consist of Single Nucleotide Polymorphisms (SNPs). SNPs are point mutations that show variation across a population and can result in a variety of effects, such as modifying transcription binding sites or changing a protein's amino acid composition. A typical representation of a SNP is to encode the genotype into integers, for example $AA = 0,\ Aa = 1,\ aa = 2$.

The grammar used to select features and parameter values for the GESVM can be expressed in Backus-Naur Form as follows:

$$\begin{aligned} N &= \{C_1,\ \gamma_1,\ C_2,\ \gamma_2,\ X,\ L,\ E\} \\ T &= \{1 - 50\} \\ S &= \{< C_1 >< \gamma_1 >< X >< L >\} \end{aligned}$$

with P defined as

$$\begin{aligned} < C_1 >, < C_2 > &::= 5 \\ < \gamma_1 >, < \gamma_2 > &::= 2 \\ < X > &::= 1 - 50 \\ < L > &::= \ < X >< E > \\ &\quad | \ < X >< L > \\ &\quad | \ < X >< L >< L > \\ &\quad | \ < X >< L >< L >< L > \\ &\quad | \ < E > \\ < E > &::= \ < C_2 >< \gamma_2 > \end{aligned}$$

The grammar begins by selecting values for C_1 and γ_1, which are used to determine a proportion of the misclassification penalty C and kernel parameter γ, respectively. For this preliminary work the parameter values C and γ were fixed at 5 and 2, respectively, and were chosen based on an initial parameter sweep of a dataset generated using a purely epistatic model. Next, a value X is chosen, which corresponds to a specific SNP in the dataset. For now we consider only models with at most 50 potential SNPs from which to select and SNPs are not allowed to be selected multiple times for an individual SVM. After selecting the initial SNP, the grammar will replace L with the appropriate line, allowing the possibility for additional SNP loci to be selected. Once the grammar reaches an E, the remaining proportion of C and γ are selected and the mapping process ends. The distribution for the expected number of SNPs selected in a randomly initialized population using this grammar is shown in Figure 1.

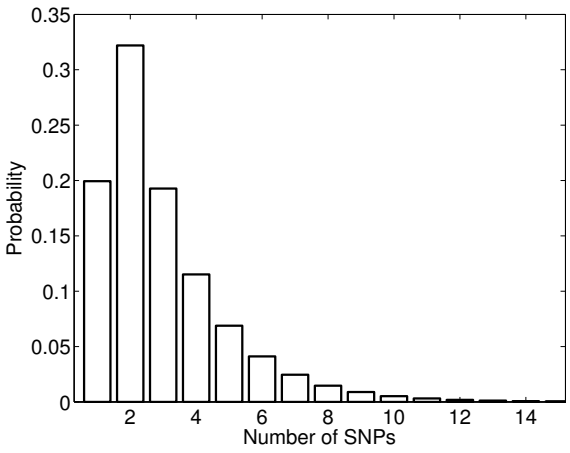

Figure 1: Distribution for the number of SNPs selected by the grammar mapping process for a population of 10^6 randomly generated binary strings.

The reason we decided to split the values of C and γ into two proportions is to allow the 'ripple effect' to influence these parameters in future analyses where they will no longer be constrained to specific values. The 'ripple effect' is a property of grammatical evolution that occurs when a mutation upstream in the binary string genome results in a different mapping sequence after that point. For example, if a mutation caused the grammar to replace L with $< X >< L >$ rather than $< E >$, then the codons that would have been used to choose values for C_2 and γ_2 are

Table 1: GE settings

population size	50
max generations	20
crossover rate	0.8
mutation rate	0.05
codon size	8
wrapper count	2
chromosome size	25
selection	tournament
fitness	classification

Table 2: Multilocus penetrance function, model M_1

	BB	Bb	bb
AA	0.11	0.1	0
Aa	0.1	0.11	0
aa	0	0	0

instead used to select an additional SNP and select another line from L. This 'ripple effect' can cause SNPs to be added or removed from the model and may shift the location in the genome where the second proportion of C and γ are obtained.

Many of the GE parameters used in this initial work are shown in Table 1. Initial populations were randomly generated, however, the codons used to select the first SNP value were modified to ensure each SNP was represented in the first generation of the population. The chromosome size was fixed at 25 codons and a one-point crossover was used. Selection was performed by keeping the proportion of the population that had the best fitness, determined by classification accuracy. Ten-fold cross-validation was used to reduce overfitting by splitting the data into different training and testing sets.

2.4 Data Generation

For the purposes of this initial work, we generated genetic models that were not purely epistatic. While the eventual goal of this work is to detect gene-gene interaction, initial evaluation of this method is made easier by using models with main effects from single SNPs in addition to epistatic effects. We used penetrance functions to define the probability of disease given a particular genotype. Two models were tested, both with similar penetrance functions but different genotype frequencies. The penetrance functions for the two models are shown in Tables 2 and 3. Minor allele frequencies were adjusted so that the heritability, or proportion of the trait (disease) variance that is due to genotype, would be modest and were calculated for both models to be $\sim$0.05 according to Culverhouse *et al.* [4] when the genotypes were generated according to Hardy-Weinberg proportions. The minor allele frequencies for model M_1 were $p(a) = 0.5$ and $p(b) = 0.5$, while for model M_2 they were $p(a) = 0.15$ and $p(b) = 0.32$. Twenty-five datasets consisting of 300 cases and 300 controls were generated for each model using the software GenomeSIMLA [5, 6] with the disease loci $A, a = \text{SNP}_1$ and $B, b = \text{SNP}_2$.

2.5 Simulation platform

All code was written in Matlab. Training of the SVM was computed using the quadratic programming routine with the `interior-point-convex` algorithm. Code is available from the authors upon request. Computations were performed using the High Performance Computing hardware at NCSU.

Table 3: Multilocus penetrance function, model M_2

	BB	Bb	bb
AA	0.3	0.3	0
Aa	0.3	0.3	0
aa	0	0	0

3. RESULTS

GESVM was applied to the case/control datasets generated from disease model M_1. The architectures of the SVMs for the most fit individuals from each cross-validation step for all 25 datasets were recorded. In this preliminary work C and γ were kept constant so only the number of SNPs and SNP values were different among individual SVM architectures. The accuracy of the best fit individuals ranged from 65.0% to 90.0% with mean 75.6% and standard deviation 4.6%. The expected accuracy of model M_1 was 73.3%. A distribution of which SNP values were included in the SVM models is shown in Figure 2. Of the best fit individuals, 6.6% included SNP_1 but not SNP_2, 7.2% included SNP_2 but not SNP_1, and 75.1% included both SNP_1 and SNP_2. A second set of classification was performed using SVM with all 50 SNPs as features. The accuracy when using all features ranged form 40.0% to 76.7% with mean 60.4% and standard deviation 6.2%.

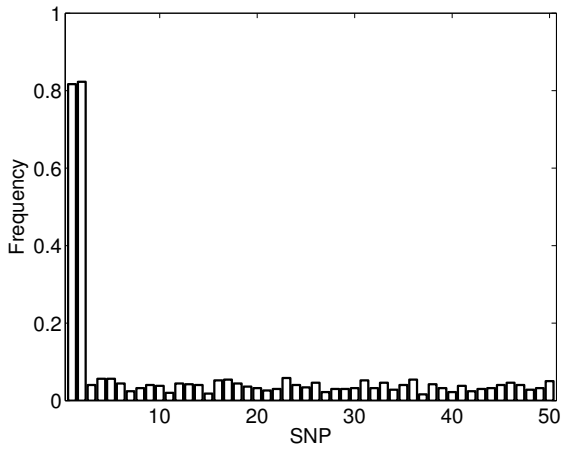

Figure 2: Frequency of SNPs included in top-performing SVMs for model M_1.

This analysis was repeated using the datasets generated from disease model M_2. The accuracy of the best fit individuals ranged from 61.7% to 78.3% with mean 69.2% and standard deviation 3.1%. The expected accuracy of model M_2 was 58.3%. A distribution of which SNP values were included in the SVM models shown in Figure 3. Of the best fit individuals, 4.5% included SNP_1 but not SNP_2, 25.1% included SNP_2 but not SNP_1, and 1.7% included both SNP_1 and SNP_2. A second set of classification was performed using SVM with all 50 SNPs as features. The accuracy when using all features ranged form 38.3% to 71.7% with mean 53.3% and standard deviation 6.5%.

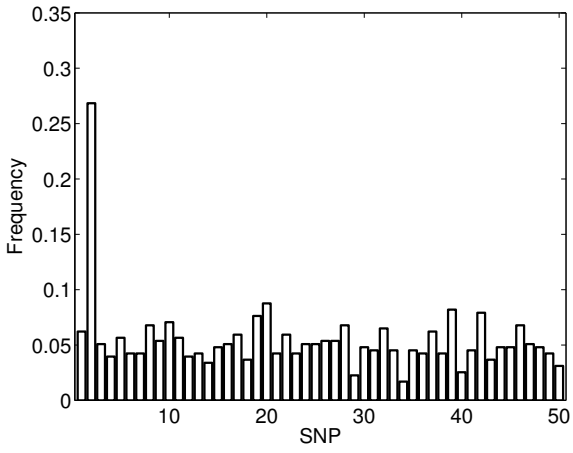

Figure 3: Frequency of SNPs included in top-performing SVMs for model M_2.

4. DISCUSSION AND FUTURE WORK

These preliminary results show promise for using GESVM as a potential method for identifying human disease susceptibility genes. The results presented in the previous section were obtained using unoptimized GE and SVM parameters, yet the best fit SVMs were still able to identify SNP_1 and SNP_2 for disease model M_1 and SNP_2 for disease model M_2. The overall decrease in ability for GESVM to identify disease associated SNPs in model M_2 compared to M_1 can be attributed to the fact that the minor allele frequencies were lower in model M_2. This caused the datasets for disease model M_2 to have a larger proportion of the generated population with haplotypes where the value of the penetrance function is nonzero. Having a larger proportion of the population consisting of incomplete penetrance haplotypes reduced the chance of finding effects from either SNP_1 or SNP_2. In a similar manner, a smaller minor allele frequency for SNP_1 relative to SNP_2 in model M_2 further reduced the ability for GESVM to detect effects from SNP_1, which resulted in GESVM only finding SNP_2.

The true power of GESVM cannot be determined from these initial studies. The next step will be to perform parameter sweeps for both GE and SVM parameters for many different datasets. These sweeps will include the GE parameters population size, number of generations, crossover rate and mutation rate; and SVM parameters C and γ. Each parameter sweep will be applied to a variety of datasets, which will have several different minor allele frequencies, heritability and sample sizes. Once this has been accomplished, application to real data sets will be studied. Only after such thorough investigation can the true potential of GESVM be discussed.

5. ACKNOWLEDGMENTS

This work was supported by National Institute of Environmental Health Sciences training grant 5T32ES007329-12.

6. REFERENCES

[1] E. Capriotti, R. Calabrese, and R. Casadio. Predicting the insurgence of human genetic diseases associated to single point protein mutations with support vector machines and evolutionary information. *Bioinformatics*, 22(22):2729–2734, 2006.

[2] S.-H. Chen, J. Sun, L. Dimitrov, A. R. Turner, T. S. Adams, D. A. Meyers, B.-L. Chang, S. L. Zheng, H. GrÃűnberg, J. Xu, and F.-C. Hsu. A support vector machine approach for detecting gene-gene interaction. *Genetic Epidemiology*, 32(2):152–167, 2008.

[3] H. J. Cordell. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, 11(20):2463–2468, 2002.

[4] R. Culverhouse, B. K. Suarez, J. Lin, and T. Reich. A perspective on epistasis: Limits of models displaying no main effect. *The American Journal of Human Genetics*, 70(2):461 – 471, 2002.

[5] S. M. Dudek, A. A. Motsinger, D. R. Velez, S. M. Williams, and M. D. Ritchie. Data simulation software for whole-genome association and other studies in human genetics. In *Pacific Symposium on Biocomputing*, pages 499–510, 2006.

[6] T. L. Edwards, W. S. Bush, S. D. Turner, S. M. Dudek, E. S. Torstenson, M. Schmidt, E. Martin, and M. D. Ritchie. Generating linkage disequilibrium patterns in data simulations using genomesimla. In *Proceedings of the 6th European conference on Evolutionary computation, machine learning and data mining in bioinformatics*, EvoBIO'08, pages 24–35, Berlin, Heidelberg, 2008. Springer-Verlag.

[7] L. Jack and A. Nandi. Fault detection using support vector machines and artificial neural networks, augmented by genetic algorithms. *Mechanical Systems and Signal Processing*, 16(2 - 3):373 – 390, 2002.

[8] T. A. Manolio. Genomewide association studies and assessment of the risk of disease. *New England Journal of Medicine*, 363(2):166–176, 2010.

[9] A. A. Motsinger-Reif, S. M. Dudek, L. W. Hahn, and M. D. Ritchie. Comparison of approaches for machine-learning optimization of neural networks for detecting gene-gene interactions in genetic epidemiology. *Genetic Epidemiology*, 32(4):325–340, 2008.

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

[11] Y. Saeys, I. Inza, and P. LarraŰaga. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517, 2007.

[12] B. Samanta, K. Al-Balushi, and S. Al-Araimi. Artificial neural networks and support vector machines with genetic algorithm for bearing fault detection. *Engineering Applications of Artificial Intelligence*, 16(7 - 8):657 – 665, 2003.

[13] V. Vapnik. *The nature of statistical learning theory*. Springer-Verlag New York Inc, 2000.

[14] Z. Wei, K. Wang, H.-Q. Qu, H. Zhang, J. Bradfield, C. Kim, E. Frackleton, C. Hou, J. T. Glessner, R. Chiavacci, C. Stanley, D. Monos, S. F. A. Grant, C. Polychronakos, and H. Hakonarson. From disease association to risk assessment: An optimistic view from genome-wide association studies on type 1 diabetes. *PLoS Genet*, 5(10):e1000678, 10 2009.