A Balanced Accuracy Fitness Function Leads to Robust Analysis using Grammatical Evolution Neural Networks in the Case of Class Imbalance

Nicholas E. Hardison
Bioinformatics Research Ctr.
Department of Statistics
North Carolina State University
Raleigh, NC 27606
nhardis@ncsu.edu

Theresa J. Fanelli
Ctr. for Human Genetics Research
Department of Molecular Physiology & Biophysics; Vanderbilt University
Nashville, TN 37232
tjf5004@psu.edu

Scott M. Dudek
Ctr. for Human Genetics Research
Department of Molecular Physiology & Biophysics; Vanderbilt University
Nashville, TN 27232
dudek@chgr.mc.vanderbilt.edu

David M. Reif
National Ctr. for Computational Toxicology; U.S. Environmental Protection Agency
RTP, NC 27711
reif.david@epa.gov

Marylyn D. Ritchie
Ctr. for Human Genetics Research
Department of Molecular Physiology & Biophysics; Vanderbilt University
Nashville, TN 37232
ritchie@chgr.mc.vanderbilt.edu

Alison A. Motsinger-Reif
Bioinformatics Research Ctr.
Department of Statistics
North Carolina State University
Raleigh, NC 27606
motsinger@stat.ncsu.edu

ABSTRACT

Grammatical Evolution Neural Networks (GENN) is a computational method designed to detect gene-gene interactions in genetic epidemiology, but has so far only been evaluated in situations with balanced numbers of cases and controls. Real data, however, rarely has such perfectly balanced classes. In the current study, we test the power of GENN to detect interactions in data with a range of class imbalance using two fitness functions (classification error and balanced error), as well as data re-sampling. We show that when using classification error, class imbalance greatly decreases the power of GENN. Re-sampling methods demonstrated improved power, but using balanced accuracy resulted in the highest power. Based on the results of this study, balanced error has replaced classification error in the GENN algorithm.

Categories and Subject Descriptors

Genetics-Based Machine Learning and Learning Classifier Systems.

General Terms

Algorithms

1. INTRODUCTION

Grammatical Evolution Neural Networks (GENN) uses grammatical evolution to evolve neural networks to detect gene-gene interactions in studies of complex human diseases [1]. GENN has shown initial successes in both real and simulated data, and while these results are encouraging, previous simulation studies have used datasets with balanced numbers of cases and controls. Unfortunately, when using standard classification error as the fitness function, many machine learning methods are not robust to class imbalance.

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

To try to solve this problem, investigators have tried techniques such as re-sampling [2] or altering the fitness metric. One metric that has been shown to be highly successful is balanced error/accuracy [3]. This metric has been shown to solve the class imbalance problem for another approach designed to detect epistasis–Multifactor Dimensionality Reduction (MDR) [4].

We assessed the performance of GENN on data with varying levels of class imbalance and show that the power of GENN using classification error decreases as the control:case ratio departs from unity. We compared three methods for addressing this concern: re-sampling methods (over- and under-sampling) and balanced accuracy as a fitness function.

2. METHODS

2.1 Grammatical Evolution Neural Networks

The steps of GENN have been previously described in detail [1]. For the purposes of the current study, an option was added to the configuration file to specify the fitness function used: classification error (CE) or balanced error (BE). BE is the inverse of balanced accuracy, defined as the mean of sensitivity and specificity [3]:

$$\text{Balanced Accuracy} = (\text{sensitivity} + \text{specificity})/2 = \tfrac{1}{2}\left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right]$$

where TP represents true positives, TN represents true negatives, FP represents false positives, and FN represents false negatives. This formula equally weights the errors within each class. In the case of balanced data, this is equivalent to standard CE.

2.2 Data Simulation

The intention of the data simulations for this power study was to mimic gene-gene interaction, or epistasis, in case-control genetic data to evaluate GENN using penetrance functions. Penetrance defines the probability of disease given a particular genotype combination by modeling the relationship between genetic

variations and disease risk. We used two well-described purely epistatic models, where the heritability (the proportion of trait variance due to genetics) ~5%. The first is referred to as the XOR model, and the second is referred to as the ZZ model [5]. Both are nonlinear models with no marginal main effects. Software described by Moore *et al* [5] was used to simulate the data.

For both models, we simulated data with a range of control:case ratios and sample sizes. For the first set of simulations, the total number of individuals in the dataset was held constant, at two different total sample sizes: 600 and 1200. For each sample size, three control:case ratios were simulated: 1:1, 2:1, and 4:1. To ensure the results seen were due to class imbalance instead of decreasing numbers of cases, a second set of simulations was done, holding the number of cases constant at 300 and 600. Again, for each number of cases, three control:case ratios were simulated. For each set of parameters, 100 replicates were simulated. Each dataset had a total of 100 SNPs, two of which were functional in predicting disease. For the models with imbalanced control:case ratios, re-sampling was performed. In the case of under-sampling (US), controls were randomly removed until a ratio of 1:1 was achieved. In the case of over-sampling (OS), cases were randomly re-sampled until a 1:1 ratio was achieved.

2.3 Data Analysis

GENN was used to analyze all epistasis models with classification error, balanced error, or classification error in combination with data re-sampling. Parameter settings remained identical between the analyses and included: 4 demes, migration every 25 generations, population size of 200 per deme, 400 generations, crossover rate of 0.9, and a reproduction rate of 0.1. Power for all analyses is reported as the number of times GENN correctly identified the correct loci with no false positives over 100 datasets.

3. RESULTS

Tables 1 and 2 show the results for all analyses, with several apparent trends. Using classification error (CE), increased imbalanced ratios greatly decreases the power of GENN. The power of GENN greatly improves when OS is used. With US, a marked decrease in power in smaller datasets with large class imbalance is seen. This trend is ameliorated somewhat in larger datasets, as well as the datasets with fixed numbers of cases. Most significantly, for all models analyzed, power recovers completely when using balanced error (BE).

Table 1. Results for constant sample size simulations for different control:case ratios (CCR).

Total Samples	CCR	XOR Power (%)				ZZ Power (%)			
		CE	BE	US	OS	CE	BE	US	OS
600	1:1	100	100	100	100	100	100	100	100
	2:1	74	100	87	98	100	100	96	100
	4:1	3	100	62	97	63	99	74	99
1200	1:1	99	100	100	100	100	100	100	100
	2:1	88	100	98	99	100	100	99	100
	4:1	6	100	85	99	59	100	94	100

Table 2. Results for constant case number simulations.

Case Count	CCR	XOR Power (%)				ZZ Power (%)			
		CE	BE	US	OS	CE	BE	US	OS
300	1:1	100	100	100	100	100	100	99	100
	2:1	80	100	91	96	100	100	99	99
	4:1	2	100	86	95	49	100	90	96
600	1:1	99	100	100	100	100	100	100	100
	2:1	83	99	93	99	100	100	98	99
	4:1	3	100	71	98	36	100	92	97

4. DISCUSSION

From these results, we conclude that balanced error should be used as the fitness metric in GENN instead of classification error, as it outperforms standard classification error and re-sampling methods. Additionally, since balanced error and classification error are mathematically equivalent in when data is balanced, there is no disadvantage to using balanced error in balanced data.

5. ACKNOWLEDGMENTS

This work was supported by National Institutes of Health grants HL65962, GM62758, and AG20135. We would also like to thank Jason H. Moore and Digna R. Velez for helpful discussions on class imbalance. This paper has been reviewed and approved for publication according to US EPA policy but does not necessarily represent the views of the Agency.

6. REFERENCES

[1] Motsinger-Reif A.A., Dudek S.M., Hahn L.W., Ritchie M.D. Comparison of Approaches for Machine Learning Optimization of Neural Networks for Detecting Gene-Gene Interactions in Genetic Epidemiology. *Genet. Epidemiol., Epub ahead of print.* (2008)

[2] Japkowicz N., Stephen S. The Class Imbalance Problem: A Systematic Study. *Intelligent Data Analysis Journal., 6* (2002) 429-450.

[3] Powers R., Goldszmidt M., Cohen I. *Short term performance forecasting in enterprise systems.* Hewlett-Packard Development Company Technical Reports, Computer Science Department, Stanford University, Stanford, CA. (2005)

[4] Velez D., White B.W., Motsinger A.A., Bush W.S., Ritchie M.D., Moore J.H. A Balanced Accuracy Metric for Epistasis Modeling in Imbalanced Datasets using Multifactor Dimensionality Reduction. *Genet. Epidemiol. 4* (2007) 306-15.

[5] Moore, J., Hahn, L., Ritchie, M., Thornton, T., White, B. Application of genetic algorithms to the discovery of complex models for simulation studies in human genetics. In *Proceedings of the Genetic and Evolutionary Computation Conference*. (New York, USA, July 9-13, 2002). Morgan Kaufman, San Francisco, CA, 2002, 1150-1155.