Feature Selection and Classification in Noisy Epistatic Problems using a Hybrid Evolutionary Approach

Drew DeHaas[1]
Drew.DeHaas@uvm.edu

Jesse Craig[1]
JesseCraig@alumni.uvm.edu

Colin Rickert[1]
Colin.Rickert@uvm.edu

Paul Haake[1]
Paul.Haake@uvm.edu

Kirsten Stor[2]
Kirsten.Stor@uvm.edu

Margaret J. Eppstein[1]
Maggie.Eppstein@uvm.edu

ABSTRACT

A hybrid evolutionary approach is proposed for the combined problem of feature selection (using a genetic algorithm with Intersection/Union recombination and a fitness function based on a counter-propagation artificial neural network) and subsequent classifier construction (using strongly-typed genetic programming), for use in nonlinear association studies with relatively large potential feature sets and noisy class data. The method was tested using synthetic data with various degrees of injected noise, based on a proposed mental health database. Results show the algorithm has good potential for feature selection, classification and function characterization.

Categories & Subject Descriptors: I.2.6 [**Learning**] Knowledge Acquisition, Parameter learning

General Terms: Algorithms, Experimentation, Performance

1. INTRODUCTION

The Vermont Department of Public Psychiatry is developing a database of mental health and substance abuse patients, comprising features ranging from general data to specific symptoms and treatments to patient outcomes. It would be useful to have a classifier for this future database that provides the functional form of interaction between features and classes. The large number of features relative to sample size requires feature selection prior to classifier construction. However, strong nonlinearities in the interactions between features preclude the use of a constructive feature selector, since individual features have little or no marginal effects.

2. METHODS

We perform feature selection with a genetic algorithm (GA), where 95% of the time recombination is performed using set intersection for set size reduction and 5% of the time with set union for repair. In preliminary testing, this intersection/union recombination operator far out-performed single-point crossover and these percentages appeared optimal. The GA uses a noisy fitness function based on classification error of a counter-propagation artificial neural network (ANN). Classifier construction is then performed on the reduced feature set with strongly typed genetic programming (GP) and, for comparison, a newly trained counter-propagation ANN.

Candidate solutions for the GA feature selector are represented by binary chromosomes of length m, where a 1 in column i means that the i^{th} feature is selected, and m is the total number of features. We size the GA population such that approximately q individuals in the population are expected to be supersets of the correct feature set of size f. With uniform binary initialization, q = 15, and f = 3, this requires a population of 120 chromosomes.

The reduced feature set is passed to a GP, which it uses to initialize a population of function trees. We use a strongly typed GP where each tree returns a boolean value at its root, representing the predicted class as a function of the features. Our GP population size is 800, based on preliminary empirical testing. For comparison, we also retrained an ANN classifier using only the reduced feature set.

We tested our method on synthetic data modeled after the proposed mental health database, generated in three steps: 1) generate 56 integer-valued and 54 boolean-valued features for each of 2000 simulated patients, 2) classify each patient based on a randomly generated nonlinear function of three features, making sure that the ratio of cases:controls is between 3:1 and 1:3., and 3) inject up to 50% noise into the classification data. For each of six noise levels, and each classification strategy (GA+ANN→GP, GA+ANN→ANN), we ran 4 repetitions on each of 10 randomly generated problems, using 40% of simulated cases and controls for training and 10% for testing. For each problem and each classifier type, we selected the best of the 4 repetitions, based on classification accuracy on the testing set. The resulting classifiers were then validated using the remaining 50% of patients.

3. RESULTS

The feature selector performed well; even at 20% noise it found all the correct features in 9 out of 10 trials and included at most 2 extraneous features. The GP classifier achieved nearly optimal results given each noise level, and increasingly out-performed the ANN as noise increased (Figure 1). When presented the selected features, the GP always reconstructed the underlying nonlinear classifier expression, even with up to 20% noise in the class data.

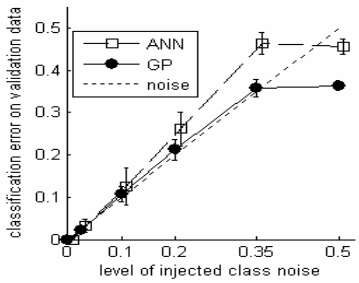

Figure 1: Comparison of classification errors of GP vs ANN

GECCO'07, July 7–11, 2007, London, England, United Kingdom
ACM 978-1-59593-697-4/07/0007.

[1] Dept. of Computer Science, and [2]Dept. of Mathematics,
University of Vermont, Burlington, VT 05405