NUANCE: A New Genetic Programming Environment for Linear and Nonlinear Equation Modeling

Geoff Hollis & Chris Westbury, Department of Psychology, University of Alberta

Abstract

In many domains of psychology, the number of variables influencing a single psychological phenomenon can be quite large. Furthermore, the relation between those variables and some measure of the phenomenon may be radically nonlinear. In such cases, the constraints imposed by conventional regression techniques can be quite limiting. However, there are other tools for function approximation that succeed at describing variable relationships where conventional regression methods fail. We introduce one such tool: the **Naturalistic University of Alberta Nonlinear Correlation Explorer** (NUANCE). NUANCE is a freely-available Java program that uses genetic programming (programming by natural selection) to search the space of relations between any number of predictors and a single value to be predicted. We explain how NUANCE works, and present example analyses in which NUANCE has outperformed the ability of conventional statistical techniques at capturing variable relationships.

How Does Genetic Programming Work?

Genetic Programming (GP) is analogous to selective breeding in biology. One way a GP system could work, and the way NUANCE works, is by making a random population of mathematical equations, deciding which equations are best fit to solve a specific problem, and then composing new equations with random parts of the parent equations. Across several generations of this selective breeding,, average and best fitness of the population members increases because each successive generation will contain more population members who have inherited useful sub-equations suited for solving the problem at hand, all chained together in a useful way.

The embodiment of a GP population member can vary from GP system to GP system. The general principles of GP work for evolving mathematical equations, decision trees, vectors and arrays, machine code, and neural networks, to name a few architectures it works with.

Experiment 1

Westbury, Buchanan, Sanderson, Rhemtulla & Phillips (2003) used GP to model how orthographic frequency and orthographic neighbourhood size impinge on lexical decision reaction times (LDRT) for words. Their best equation estimated LDRTs with $r=0.48$ ($p < 0.0001$) on a 450-item dataset used to evolve the equations. The equation's estimates generalized to a 150-item dataset not used for evolution ($r=0.61$, $p < 0.0001$). A linear regression equation correlates with the 450-item dataset at $r=0.22$ ($p < 0.01$), and generalizes to the unseen data set with $r=0.20$ ($p < 0.01$). We replicated the Westbury *et al* work with NUANCE to see how it compares in power to a similar program.

Method

We ran NUANCE ten times with the 450-item dataset used by Westbury *et al*. We used a population size of 2500 and let the program run for 100 generations, using the averaged multitest fitness function built into NUANCE (see Westbury *et al*, 2003, for the mechanics of averaged multitest fitness).

Results

NUANCE performed similarly to Westbury *et al*'s program ($r=0.51$ for the 450-item dataset, $r=0.60$ for the 150-item dataset). The predictions of the two equations were identical, with the exception of estimates at the ends of the RT ranges ($r=0.98$ for the 450-item dataset and $r=0.97$ on the 150-item dataset). All correlations were significant at $p < 0.0001$.

NUANCE estimate versus prior estimate

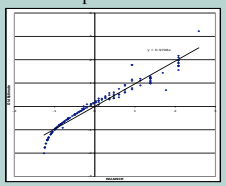

Figure 1. Comparing Westbury *et al*'s equation with the NUANCE-evolved equation. Functionally, both are near-identical.

Experiment 2

Nonlinear relationships can often be hard to discern by simply looking at the data, and there are occasions when conventional statistical tools can underestimate how strong the relationship between some set of variables actually might be. We wanted to know if NUANCE can find functions to describe such relationships better than conventional statistical tools.

Method

We obtained a dataset from the Data And Story Library (DASL, at http://lib.stat.cmu.edu/DASŁ) that presents the relationship between average cigarette consumption per capita and leukemia deaths per 100,000 people in 44 American states for 1960. We used linear regression, cubic regression, and NUANCE to model the relationship between average cigarette consumption and leukemia deaths. The settings we used with NUANCE were the same as those in experiment 1, except we used an age-weighted fitness function instead of an averaged multitest fitness function (the differences are discussed in NUANCE's manual).

Results

The linear correlation between cigarette consumption and leukemia deaths in this data set is insignificant, at $r= -0.07$ ($p > 0.05$). A cubic regression was better, but still insignificant ($r=0.32$, $p > 0.05$). NUANCE produced an equation correlating cigarette consumption with leukemia deaths with cigarette consumption at $r=0.52$ ($p < 0.0001$). Figure 2 has the NUANCE estimates superimposed on a scatter plot of the two variables.

The relation between leukemia death and cigarette consumption

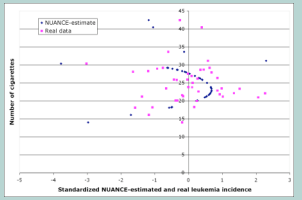

Figure 2. Predicting leukemia deaths from cigarette consumption. In this graph the NUANCE-evolved estimate and the raw leukemia data are both graphed, in standardized units, against cigarette consumption.

Conclusions

NUANCE can capture relationships that are difficult to notice. It is comparable in power to the Westbury *et al* program, much faster, and easy to use. NUANCE could benefit many people in virtually all fields of research. Because of this, **we have released NUANCE for free download at:**

http://www.ualberta.ca/~hollis/nuance.html

References

Westbury, C., Buchanan, L., Sanderson, S., Rhemtulla, M., & Phillips, L. (2003). Using genetic programming to discover non-linear variable interactions. Behavior Research Methods, Instruments, and Computers. 35:2, 202-216.