Intelligent Data Analysis in medicine and pharmacology – IDAMAP2001; September 4th, 2001 London
http://magix.fri.uni-lj.si/idamap2001/scientific.asp

Genetic programming applied to severe diseases diagnosis.

James Cunha Werner (wernerjc@sbu.ac.uk)
Terence C. Fogarty (fogarttc@sbu.ac.uk)

SCISM
South Bank University
103 Borough Road
London SE1 0AA

Abstract. This paper addresses the problem of how to obtain a mathematical discriminate function to quantify the severity of a disease with genetic programming (GP). It was applied to breast-cancer testing because it is important to develop a reliable but inexpensive test to identify women with high risk for a more expensive and accurate clinical procedure.

Introduction

Artificial intelligence can help with information extraction from databases facilitating better decision making in complex systems. One possible approach is the building of a mathematical model to allow the simulation of future events based on past records. The method consist of applying an algorithm that has data as input and a model as output satisfying some optimisation criteria such as minimum error.

The modelling algorithm.

The genetic programming (GP) algorithm (Holland 1975, Koza 1992) mimics the evolution and improvement of life through reproduction, when each individual contributes with its own genetic information to building a new individuals with greater fitness to the environment and higher chances of survival. Each 'individual' in a generation represents, with its chromosome, a feasible solution to the problem; in our case, a discriminate function to be evaluated by a fitness function.

The best individuals are continuously being selected, and crossover and mutation take place. Following a number of generations, the population converges to the solution that best represents the discrimination function.

There are two kinds of information defined for the algorithm: terminals (variable values and random numbers) and functions (mathematical functions used in the generated model).

The final GP output is a model or a function (we term discriminate function) of the process under study. This is different from other techniques like Receiver Operating Characteristic (ROC) curve analysis which search for a possible cut-off point or criterion value to discriminate between two populations (one with a disease, the other without). There will be some cases with the disease correctly classified as positive (TP = True Positive fraction), but some cases with the disease will be classified negative (FN = False Negative fraction). On the other hand, some cases without the disease will be correctly classified as negative (TN = True Negative fraction), but some cases without the disease will be classified as positive (FP = False Positive fraction) – see ROC.

Processing description.

The software we have developed is an adaptation of LilGP (see reference), where GP is structured in a pre-compiled library, with other artificial intelligence procedures (such as genetic algorithm (GA), an adaptive algorithm (AA), neural networks (NN), and fuzzy control (FC)), integrating the model obtaining process with GP from with real time adaptation by the GA, AA, NN, or FC (a second step of the work, as described in "Future works and conclusion"). This has been applied to the Wisconsin Diagnostic Breast Cancer (WDBC: see reference) using BREAST-CANCER-WISCONSIN.DATA for pre-processed data analysis, available in the Internet.

A filter reads the data and stores it in memory and a fitness function evaluates the accuracy of the discriminate function.

The discriminate function is checked against the type of tumour (benign or malign) to find the fitness (% of hit marks).

The original data is divided into 10 blocks, using each block to test while the rest are used to train the algorithms. The 10 test blocks form all the original database.

Processing output uses an Excel interface, generating a spreadsheet file ready for analysis by a technician, a computer scientist or a physician. A text format file contains the equation of discriminate function and its performance.

Breast-cancer discriminate function with pre-processed data.

The pre-processed dataset consist of the following attributes:

Attribute	GP Function name	Domain
Sample code number		id number
Clump Thickness	clthic	1 - 10
Uniformity of Cell Size	uncsiz	1 - 10
Uniformity of Cell Shape	uncsha	1 - 10
Marginal Adhesion	maradh	1 - 10
Single Epithelial Cell Size	sepcsz	1 - 10
Bare Nuclei	barnuc	1 - 10
Bland Chromatin	blachr	1 - 10
Normal Nucleoli	nornuc	1 - 10
Mitoses	mitose	1 - 10
Class		2 for benign, 4 for malignant

The GP parameters are 100 individuals and 500 generations, 60% crossover probability and 20% mutation probability. The functions are: multiplication - *, sum - +, subtraction - -, division - /, sine - sin, cosine - cos, exponential – exp.

The fitness function punishes solutions with false negative, searching for a solution with as few cases as possible. We use the fitness function coded as:

$$fitness = \frac{ok}{ok + false\_positive + punish * false\_negativ}$$

where *ok* are the correct predictions and *punish* is the punish factor (=10) for false negative.

The bibliography shows studies with several techniques (West 2000): Multilayer perceptron (MLP), general regression (GR), radial basis function (RBF), mixture of experts (MOE), LDA, logistic regression, K search neighbour, and Kernel. All these techniques have more false negatives than GP, which is very concerning in medical diagnostics. Table I shows the different results.

Discriminate function (Data block 8):accuracy 100%

```
(- (- (+ (+ (* (+ sepc barn) (+ 6.70 clth)) (- (exp
uncs) 93.31)) (+ (- (exp (- norn (/ mito 53.37))) (cos
(sin (* (/ (+ (exp (sin (+ (+ (cos (exp mito)) (sin barn))
(- mara sepc)))) clth) (exp 93.31)) (/ (cos (exp (* (+
(cos (cos (* (exp (cos (exp mito))) (/ (* (/ norn mito)
(sin (cos (+ sepc barn)))) (* (sin norn) (/ (* norn norn)
clth)))))) (- (* blac (/ mara norn)) (/ mara norn))) (cos
(exp blac))))) (- blac blac)))))) (sin (* (+ (cos 91.34) (/
(* (cos mara) (* (* (+ sepc barn) (exp blac)) norn)) (*
(sin norn) (/ (* norn norn) clth)))) (- mara (+ sepc
barn)))))) (sin (sin (* (/ norn norn) (- mara sepc)))))
(sin (exp (cos (/ (* norn norn) blac)))))
```

Tab. I Different approach to cancer diagnostic, with genetic programming solution (GP).

Method	OK (%)	% False negative	% False positive
MLP	**0.957206**	**0.087448**	**0.018594**
GR	**0.967647**	**0.054393**	**0.020408**
RBF	**0.970441**	**0.030126**	**0.029252**
MOE	**0.962941**	**0.062762**	**0.023129**
LDA	**0.9633968**	**0.0711297**	**0.018018**
Logistic	**0.972182**	**0.029289**	**0.027027**
K neighbour	**0.967789**	**0.033473**	**0.031532**
Kernel	**0.95022**	**0.117155**	**0.013514**
Method	OK (%)	% False negative	% False positive
GP – Test average	**0.963235**	**0.008368**	**0.05180**

With genetic programming it is possible to search for a solution with a specific shape, such as less false negatives with the concomitant cost of a decrease in false positive the accuracy.

Future works and conclusion.

Through the use of GP we show that it is possible to find a discriminate function that separate healthy people from possibly severely diseased patients.

The future work is obtaining the optimal therapy, due the application of Genetic Control heuristics (Werner (1999)), with the division of the challenge into the steps:

1. GP used to obtain a model of the process from historic data (discriminate function in this paper).

2. GP to obtain the optimal control policy (in the breast-cancer case, a gene therapy as in Hanania(1995) or Watanabe(2000)) or GA to optimise parameters of radiation therapy treatment plans (Chambers (2000)), to maximise the performance index (minimising shifts in clinical data) taking the dis criminate in secure values.

3. Differences of the real body system needs adaptation of the optimal control policy obtained in step 2. This would involve applying step 2 again, closing the loops.

We are looking for a medical centre with oncology skills and more complete database available for a partnership, where the aim will be complete this job applying Genetic Control heuristic, driving a solution to test and treat severe diseases.

The authors acknowledge FAPESP/Brazil for sponsoring the PhD research and CNPq/Brazil for granting a doctoral scholarship.

References.

CHAMBERS,L.; *"The practical handbook of Genetic Algorithms"* Chapman & Hall/CRC,2000.

Hanania,.G. & all; *"Recent advances in the application of gene therapy to human disease*" The American journal of medicine, vol.99, November 1995, pag. 537.

HOLLAND,J.H. *"Adaptation in natural and artificial systems: na introductory analysis with applications to biology, control and artificial intelligence."* Cambridge: Cambridge press 1992 reedição 1975.

KOZA,J.R. *"Genetic programming: On the programming of computers by means of natural selection."* Cambridge,Mass.: MIT Press, 1992.

LilGP *"Genetic Algorithms Research and Applications Group (GARAGe)"*, Michigan State University; http://garage.cps.msu.edu/software/lil-gp/lilgp-index.html

ROC Receiver Operating Characteristic (ROC) curve analysis http://www.medcalc.be/roccman.html

WDBC Dr. William H. Wolberg, General Surgery Dept., wolberg@eagle.surgery.wisc.edu; W. Nick Street, Computer Sciences Dept., street@cs.wisc.edu; Olvi L. Mangasari an, Computer Sciences Dept., olvi@cs.wisc.edu ; University of Wisconsin, 1210 West Dayton St., Madison, WI 53706 http://www.ics.uci.edu/~mlearn/MLRepo sitory.html

Watanabe,T.; Sullenger,B.A.; "*RNA repair: a novel approach to gene therapy*" Advanced drug delivery reviews 44(2000) 109-118.

Werner,J.C.; *"Active noise control in ducts using genetic algorithm*" PhD. Thesis - São Paulo University- São Paulo-Brazil-1999.

West,D; West,V; "Model selection for a medical diagnostic decision support system: a breast cancer detection case" Artificial Intelligence in medicine 20(2000)183-204.