Genetic programming applied to Collagen disease & thrombosis.

James Cunha Werner, Terence C. Fogarty

SCISM
South Bank University
103 Borough Road
London SE1 0AA
{werner,fogarttc}@sbu.ac.uk

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 thrombosis testing because it is important to develop a fast, reliable and accurate test to identify the mechanism of thrombosis occurrence.

1 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 optimization criteria such as minimum error.

2 The modelling algorithm

The genetic programming 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 (variables values and random numbers) and functions (mathematical functions used at generated model).

3 Processing description

The software we have developed is an adaptation of LilGP (see reference), where genetic programming (GP) is structured in a pre-compiled library, with others artificial intelligence procedures (like genetic algorithm (GA), adaptive algorithm (AA), neural network algorithm (NN), and fuzzy control algorithm (FC)), integrating model obtaining process with GP (termed Cognitive Structure) from real time adaptation by GA, AA, NN, or FC (a second step of the work, as described in "Future works and conclusion"). It has been applied to the database collected at Chiba University hospital (see DC2001). Each patient came to the outpatient clinic of the hospital on collagen diseases, as recommended by a home doctor or a general physician in the local hospital.

Collagen diseases are auto-immune diseases. Patients generate antibodies attacking their own bodies. For example, if a patient generates antibodies in lungs, he/she will chronically lose the respiratory function and finally lose life. The disease mechanisms are only partially known and their classification is still fuzzy. Some patients may generate many kinds of antibodies and their manifestations may include all the characteristics of collagen diseases.

In collagen diseases, thrombosis is one of the most important and severe complications, one of the major causes of death. Thrombosis is an increased coagulation of blood, that cloggs blood vessels. Usually it will last several hours and can repeat over time. Thrombosis can arise from different collagen diseases. It has been found that this complication is closely related to anti-cardiolipin antibodies. This was discovered by physicians, one of whom donated the datasets for discovery challenge. Thrombosis must be treated as an emergency. It is important to detect and predict the possibilities of its occurrence.

3.1 Genetic Programming Training

A filter reads all tables available, and update a relation table with lab_exam (57542 records), spc_exam (801 records), and patient_info (1239 records) pointers which same identification and data, reducing the number of records to 262 (231 none, 1 mild, 11 severe and 18 most severe). Each field suffers the following substitution: '-' by –1.0, '+' by 1.0 and '+-' by 0.0, omitted values by field average value, and range values by its boundary (for example <2.0 by 2.0).

The discriminate function accuracy is checked against the diagnostic (negative, most severe, severe or mild) to find the fitness (% of hit marks).

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

Two different approach are adopted to determinate the discriminate function:

1. Using all data available in relation tables: GOT,GPT,LDH, ALP,TP,ALB, UA,UN,CRE, TBIL,TCHO,TG, CPK,GLU,WBC, RBC,HGB, HCT,PLT,PT, APTT,FG,PIC,TAT, TAT2,UPRO, IGG,IGA,IGM, CRP,RA,RF,C3, C4,RNP,SM, SC170,SSA,SSB, CENTROMEA, DNA,DNA-II, ACLIGG,ACLIGM, ANA,ACLIGA,KCT, RVVT,LAC, time evolution (dt), random number. GP Functions are multiply, add, subtract, and divide. Fitness function punishes wrong solutions, searching for a solution with as few cases as possible. We use the fitness function coded as:

$$fitness = \frac{ok}{ok + punish * nok}$$

where *ok* are the correct predictions and *punish* is the punish factor (=20) for wrong.

Discriminate function:

```
   (- ACLIGM (* (* (+ (+ LAC (* TAT (- (- (+ (* (* IGM (/ LDH
LAC)) (* UA (+ (/ KCT IGG) ALB))) (* UA (/ LDH LAC))) (/ (+ (+
PT C4) C4) UN)) ACLIGM))) ANA) (+ (+ C3 (- LDH (+ UA IGG)))
LAC)) (+ (+ (* (* (* IGM  (/ LDH LAC)) (/ (/ (/ UA (* (+ (/
KCT ACLIGG) (* RF IGA)) (+ (* (/  PLT PIC) (+ LDH TCHO)) (+ (-
(* UA APTT) (* IGA TAT2)) (/ ACLIGG HGB))))) IGG) (* WBC UN)))
HCT) (/ (* (* TAT (- (/ ALP UA) IGG)) (- (- (* (/ LDH LAC) (-
TP C3)) (/ (+ (+ PT C4) C4) UN)) ACLIGM)) (* (/ IGA (- GOT
RBC)) (/ (* TAT2 HCT) (/ (/ (/ UPRO SM) (+ (+ UA (+ (+ TCHO (-
CENTROMEA LAC)) ACLIGG)) (- (* (- (* UA APTT) (* IGA TAT2))
(+ (* TAT (+ PT (+ RBC (+ UA IGG)))) TP)) (+ UA IGG)))) (+
ACLIGM (+ (* (+ (+ (+ (* (* IGM (/ LDH (+ (/ (+ RBC (/ LDH
LAC)) RF) (* UA  APTT)))) TP) CENTROMEA) (* (+ PT C4) (- (+ (/
(- LDH (+ (/ KCT ACLIGG) (* RF IGA))) (/ ACLIGA SSB)) C3)
dt))) (* (+ (* (+ DNAII IGA) HCT) (/ HCT LAC)) (+ RBC (/ (+
RBC (- (* (/ (- TG WBC) GOT) (- (+ (/ 0.08 ACLIGA) (+ HGB PT))
dt)) (/ (+ (* (* IGM (/ LDH LAC)) TP)  CENTROMEA) C3))) RF))))
HCT) (* IGG GPT))))))))) (+ (* TAT (- (+ (+ C3  (- LDH (+ UA
IGG))) (- (* C4 TAT2) LDH)) (+ UA (/ (+ (+ (/ (/ (+ SM GOT) (*
WBC UN)) (+ (/ (* (* GLU 0.03) (/ ALP UA)) RF) (* UA APTT)))
(+ (+ (- (+ RBC (+ TG (/ (+ (* (* RF IGA) HCT) C4) (+ ACLIGM
(- (+ (- TP C3) (/ C3 HGB)) (/ (- TG WBC) GOT)))))) ACLIGM) (+
TAT (+ (/ (* CENTROMEA (/ (* C4 TAT2) (/ (+ RBC (* dt ACLIGA))
(* SM SC170)))) (* (/ HGB (- (/ ALP UA) RBC)) (/ ALP UA)))
TP))) (/ (/ UPRO SM) (/ (+ RBC (* ACLIGM HGB)) GOT)))) (- (*
(- (* UA APTT) (* IGA TAT2)) (+ (* TAT (+ PT (+ RBC (+ (+ PT
(+ (+ (* (* IGM (/ LDH LAC)) RNP)  CENTROMEA) (* (/ (- TG WBC)
GOT) (- (+ (/ (/ (/ (+ DNAII IGA) (/ GPT ACLIGM)) RF) C3) dt))))
(+ (/ (+ (- (* (+ (* C4 TAT2) (- (* C4 TAT2) PT)) (+ RBC (+ (+
PT C4) (- CENTROMEA LAC)))) UA) C4) (+ ACLIGM (- (+ (-
CENTROMEA LAC) (/ LDH LAC)) (- CENTROMEA LAC)))) (* (* ACLIGM
HGB) (/ HCT LAC))))))) TP)) (+ UA IGG))) RF)))) (- (/ UPRO SM)
LAC)))))
```

2. Using only lab_exam and patient_info tables: GOT,GPT,LDH, ALP,TP,ALB, UA,UN,CRE, TBIL,TCHO,TG, CPK,GLU,WBC, RBC,HGB, HCT,PLT,PT, APTT,FG,PIC,TAT, TAT2,UPRO, IGG,IGA,IGM, CRP,RA,RF,C3, C4,RNP,SM, SC170,SSA,SSB, CENTROMEA, DNA,DNA-II, time evolution (dt), random number. GP Functions are

multiply, add, subtract, and divide. Fitness function punishes wrong solutions, searching for a solution with as few cases as possible. We use the fitness function coded as:

$$fitness = \frac{ok}{ok + punish * nok}$$

where *ok* are the correct predictions and *punish* is the punish factor (=20) for wrong.

Discriminate function:

```
   (* (- (- LDH (/ (+ IGA (- IGA GPT)) (/ LDH (- IGA GPT)))) (- (*
(- (-  (+ TP LDH) (- IGA GPT)) (- (* (/ (- IGA TCHO) (- (+ (/ (-
ALB SSB) (/ (- (- C4 (+ (- (- APTT TP) dt) PT)) LDH) (- IGA
GPT))) (+ (+ GPT TAT2) (* (- (/ (/ (- UN GLU) (* (* (* (/ LDH
RNP) RBC) dt) (+ ALB GOT))) (- (+ (- (* (* DNAII SC170) RBC) (-
(+ TAT2 GLU) LDH)) (+ (- (* TCHO CRE) (/ dt WBC)) RNP)) (* (/ (+
(/ (+ GLU (+ LDH FG)) (- IGA TCHO)) (/ dt UA)) (/ (+ GLU (+ LDH
FG)) (- IGA TCHO))) (* (- (+ CRP  GPT) (- 0.73 ALP)) (* (/ GPT
GPT) (- C3 SC170)))))) (/ (* (- (+ (+ C4 TG) (/ TG TAT)) (+ ALB
GOT)) CRE) (* (/ ALB CPK) dt))) (+ (- (- PLT (/ (- (- (- CPK (+
ALB GOT)) (/ (- IGA GPT) DNAII)) (- (+ TP LDH) (- IGA GPT))) (+
(/ (* (/ (+ (- IGA GPT) (+ RNP LDH)) (- TAT TCHO)) SC170) SSA) (/
C4 RBC)))) (+ PIC TG)) (/ (+ CRE UPRO) (+ ALB  GOT)))))) (* (/ (/
CRE CENTROMEA) (/ (- (* (/ C4 DNAII) (- IGA GPT)) LDH) (* (* (/
ALB CPK) dt) (/ LDH RNP)))) (/ (+ (- IGA GPT) (+ RNP LDH)) (- TAT
TCHO))))) dt) (* (- (- (* (/ C4 DNAII) (- IGA GPT)) LDH) (- (+ (+
TP LDH) LDH) (+ (* (* (/ ALB CPK) dt) (/ LDH RNP)) (- (- (- CPK
(+ ALB GOT)) (/ (- IGA GPT) DNAII)) (- (+ TP LDH) (- IGA
GPT)))))) RBC))) RBC) LDH)) (- C4 (+ (- (* (/ (+ LDH FG) (+ (/ (*
(/ (+ (/ (+ FG UA) SSA) (- PLT GPT)) PIC) (- (- HGB HGB) (/ (-
ALB SSB) (/ (- (* (/ C4 DNAII) (- IGA GPT)) LDH) (* (* (/ ALB
CPK) dt) (/ LDH RNP)))))) SSA) (- (+ (- IGA (- (+ (* (+ (- (+ TG
TAT2) (* HGB CENTROMEA)) (/ (/ UPRO APTT) (- HCT LDH))) (+ GLU
CRP)) (+ GLU CRP)) (+ TP LDH))) (- IGA (- (- IGA (- HCT UN)) (+
TP LDH)))) (- (- PLT (/ LDH RNP)) (/ (/ (* (+ (/ (+ CRE UPRO) (+
ALB GOT)) (/ C4 RBC)) RBC) DNA) (- PLT IGG)))))) dt) (- C4 (+ (-
(- APTT IGA) (* (- (* (* DNAII SC170) RBC) (- (* (- (- (- IGA (+
(/ TP (- SSB CRE)) SSA)) (+ TP LDH)) (- (/ (- (+ TP LDH) (- PLT
GPT)) (+ (* RNP (/ (* TAT ALB) (/ LDH RNP))) (+ LDH FG))) LDH))
dt) (+ (* TAT2 (/ LDH RNP)) (- IGA GPT)))) RBC)) PT))) (- (- (-
(/ (- DNA dt) (* (+ GPT TAT2) (- TBIL GOT))) (/ (* (/ SM CRE) (-
TBIL (* PT TP))) (+ (- SSB HGB) (+ SC170 IGA)))) (/ (- GOT 0.95)
(* TBIL CRE))) (- (- (- (+ LDH FG) (/ (* (/ SM CRE) (- TBIL GOT))
(+ (- SSB HGB) (+ SC170 IGA)))) (* (- (- IGA (- (* (* (* (/ LDH
RNP) RBC) (/ C4 DNAII)) dt) (/ (* (- (+ (/ C4 DNAII) CRE) (/ SM
CRE)) (- TP DNAII)) TG))) (/ (/ 0.65 TG) (+ IGA (- (/ (- DNA dt)
(* (/ SM CRE) (- TBIL GOT))) (* (* (+ (* (+ ALB IGM) (/ (- ALB
SSB) (/ (- (+ (- (+ (/ PT IGA) (- LDH IGA)) (- (* FG TBIL) (+ RNP
APTT))) (+ (- (* TCHO CRE) (/ dt (- (* TCHO CRE) (/ dt WBC)))) (-
(- RA HCT) (+ RBC 0.76)))) LDH) (* (- (+ (- IGA (- (- (/ (+ (/
GPT GPT) (+ GLU CRP))  (+ TAT2 GLU)) C4) (+ TP LDH))) (- PLT
IGG)) (/ (/ CRE CENTROMEA) (/ TP SSA))) (- dt HCT))))) SSA) (+ TG
TAT2)) RBC))))) RBC)) (/ (/ CRE CENTROMEA) (* DNAII SC170)))))))
```

Table 1. Accuracy results of discriminate function for thrombosis diagnostic

		Case 1		Case 2	
Population		100		100	
Generations		3924		136538	
%CrossOver		60		60	
%Mutation		20		20	
Results		# hit the mark	%	# hit the mark	%
	None	172	**74**	178	**77**
	Mild	1	**100**	1	**100**
	Severe	10	**90**	11	**100**
	Most severe	18	**100**	18	**100**

One interesting result is that the data available in lab_exam table is enough to determine the discriminate function, with the cost of too much processing evaluation. The algorithm training was done obtaining a very high accuracy discriminate function. Pay attention that the 90% case is one element fault in 11 (Table 1).

3.2 Test dataset results

Lets consider the case where Lab and antibody date exams differs into one month, i.e., for each lab month antibody exam would be in the same month, one month after or before totalising 1988 records (1564 none and 424 yes - 1 mild, 250 severe and 173 most severe). We use the case 1 program due its fast convergence.

Table 2. Case 1 test results

		Case 1	
Population		100	
Generations		3924	
%CrossOver		60	
%Mutation		20	
Results		# hit the mark	%
	None	968	**61**
	Mild	1	**100**
	Severe	231	**92**
	Most severe	168	**97**
Total		400	**94**

Table 2 presents the accuracy of 94% in thrombosis prediction of 1988 records, where only 262 where used to training and the different date would introduce some noise/perturbation into the model.

4 Future works and conclusion

Starting from lab_exam and antibody_exam results, through the application of GP we show it is possible to find a discriminate function that separate occurrences of Thrombosis, with very low false negative paying the price of false positive increase. This is a fail safe condition, because the false positive patient will be drive to more specific analysis, without risks for his health.

Through the analysis of the discriminate function it's possible to determine what are the dominate variables into the model, and with a biological analysis understand the thrombosis diseases and collagen bases and mechanisms. We are looking for medical centre with skills and more complete database available for a partnership. 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. Genetic programming to obtain a model of the process from historic data (discriminate function in this paper).
2. Genetic programming to obtain the optimal control or genetic algorithm to optimise therapy treatment plans, to maximise the performance index (minimising shifts in clinical data) taking the discriminate in secure values.
3. Differences of the real body system need adaptation of optimal control obtained in step 2. It would be due applying step 2 again, into a close loop.

The authors acknowledge FAPESP/Brazil for sponsoring the PhD research, CNPq/Brazil for granting a doctoral scholarship, and to South Bank University for our financial support.

References

DC2001 Discovery Challenge 2001 – Guide to medical data on Collagen Disease & thrombosis http://www.uncc.edu/knowledgediscovery/DataDesc.htm

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

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