Risk Evaluation using Evolvable Discriminate Function.

James Cunha Werner, Tatiana Kalganova

Department of Electronic & Computer Engineering, Brunel University
Uxbridge, Middlesex, UB8 3PH
{jamwer2000@hotmail.com, Tatiana.Kalganova@brunel.ac.uk}

Abstract. This essay proposes a new approach to risk evaluation using disease mathematical modeling. The mathematical model is an algebraic equation of the available database attributes and is used to evaluate the patient condition. If its value is greater than zero it means that the patient is ill (or in risk condition), otherwise healthy. In practice risk evaluation has been a very difficult problem mainly due its sporadic behavior (suddenly, the patient has a stroke, etc as a condition aggravation) and its database representation. The database contains, under the label of risk patient data, information of the patient condition that sometimes is in risk condition and sometimes is not, introducing errors in the algorithm training. The study was applied to Atherosclerosis database from Discovery Challenge 2003 - ECML/PKDD 2003 workshop.

Objectives

This essay addresses the problem of obtain a mathematical model able to discriminate risk patients and forecast acute conditions. The mathematical model is a function of the available attributes of the database and uses the predefined classification of risky patients (NORMAL or RISK / PATHOLOGICAL).

The main bottleneck of risk evaluation is that the patient sometimes is in a risky condition and sometimes is not, and all database records are labeled as risk condition.

Our main objective is to obtain the discriminate function able to evaluate the risk condition where sometimes a risky patient is in normal condition (where there is no indication of any risk). To obtain the discriminate function, we use genetic programming, an artificial intelligence approach which output is a mathematical model of the predefined operators and database attributes.

We apply for the workshop to discuss how to extract knowledge from our model. The comparison of our results with other different approaches and expert presentations during the conference will allow us to understand it better, and develop effective methods of knowledge extraction and representation.

A detailed discussion about many diagnostic methods, the comparison with discriminate function and its results for breast cancer and collagen disease is available in Werner and Kalganova [1] and Werner and Fogarty [2],[3].

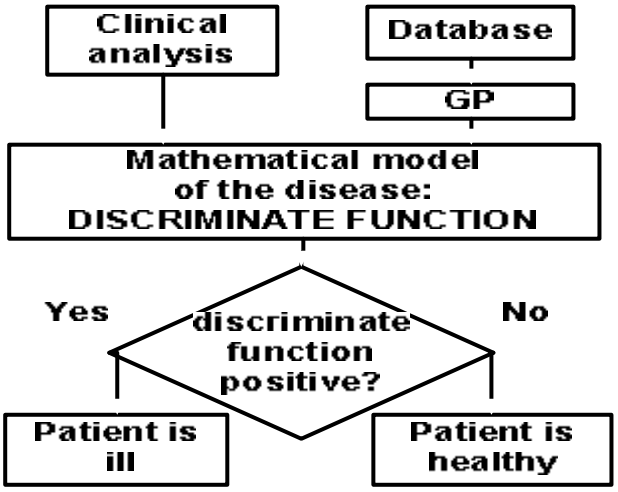
The methodology

Our approach differs from all previous approaches because it generates a mathematical algebraic model (discriminate function) used to classify the patient data. We define the operators (such as summation, subtraction, multiplication, etc) that should be used in the model assembly. Any type of model can be obtained by Genetic Programming (GP).

Discriminate function maps the original multi dimensional space in a one-dimensional real number image. The output space has a threshold with separate diagnostic classes. In this essay the origin was adopted as a threshold: positive values mean a risky patient and negative values a normal patient.

Fig. 1 shows the methodology to obtain the discriminate function. The available information from the database is used to test each different model, and the number of correct prediction is used to evaluate the quality of each model trial.

The discriminate function can be used to evaluate new patient data, classifying his condition because it contains the disease dynamics. It means that understanding the model we will be able to understand the disease mechanism, and develop new treatments. The time behavior of the discriminate function values helps to forecast possible crisis.



Genetic Programming

GP is an optimization algorithm which mimics the evolution and improvement of life through reproduction. Individual in the population represents a different algebraic equation (discriminate function). Each individual contributes with its own genetic information to the building of new ones (offspring) adapted to the environment with higher chances of surviving. This is the basis of genetic algorithms and programming [4], [5], [6], and [7].

Fig 2 shows the algorithms stages. The first stage consists in the generation of random models for each individual of the population and model accuracy is evaluated.

The evaluation process measure the performance of each model in the classification of the data available in the database. This performance (termed fitness function) is used to select the parents that contribute with their chromosomes.

Genetics operators include mutation (the change of a randomly chosen bit in the chromosome) and crossover (the exchange of randomly chosen slices between two chromosomes).

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 performs the best, e.g., the best model that represent the disease and is able to classify the patient risk.

The software we have developed is an adaptation of LilGP [8], where GP is structured in a pre-compiled library. Outputs are written in Excel XLS format direct from the program, to generate an accessible and functional Human-Computer Interface (HCI).

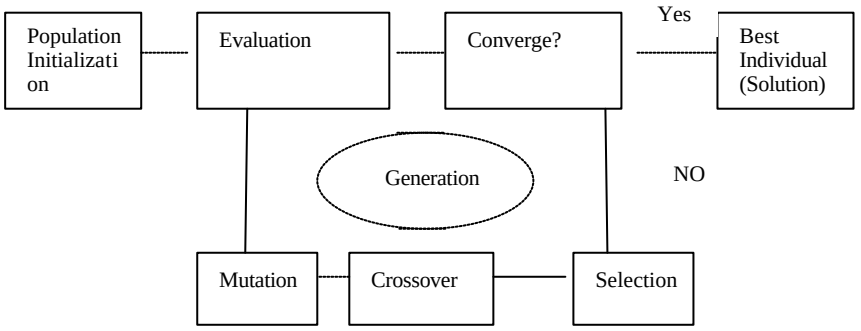


Fig. 2. Genetic programming stages

Chromosome representation. The chromosome represents the model of the problem solution using trees. A tree is a model representation that contains nodes and leaves.

Nodes are mathematical operators. We have used multiplication, addition, subtraction, and division. Leaves are terminals (the attributes of the dataset and numbers). The discriminate function in a GP context is a tree using operators (or so

called Functions) and leaves (or so called Terminals). Let us consider the following discriminate function:

$$X_1+3.14 \cdot X_2+5.3 / X_3$$

In the tree representation it can be rewritten as following:

$$(+ X_1 (+ (\cdot\ 3.14\ \ X_2) (/ 5.3\ \ X_3)))$$

where X_1, X_2, and X_3 are the attributes of the clinical data, and multiplication(·), addition(+), subtraction (-), and division(/) are the operators. Replacing the values of the clinical data in the equation results in a number which should be positive (the patient is ill) or negative (the patient is healthy).

Genetic operators. Trees are manipulated through genetic operators. The crossover operator points a tree branch and exchanges it with another branch and obtains new trees. The mutation operator changes the branch for a random new branch. The length of the chromosome is variable.

The probability of crossover is 60% and the probability of mutation is 20%. We adopt a high value of the mutation probability to spread the population over all solution space.

Fitness function. Fitness function defines the quality of chromosome as a solution to the problem. It is a numerical positive value. The dataset is used to obtain the model that maximizes the fitness function.

The fitness function F used in the disease diagnostic is the accuracy of the model:

$$F = \frac{N_{ok}}{N_{ok} + N_{Not}} \quad (1)$$

where N_{ok} is the number of correct forecast and N_{not} is the number of false forecasts.

To analyze the knowledge represented in the discriminate function, the separation between positive and negative cases and the influence of each variable, we introduced a plot of the partial derivative with respect to a variable by the difference in the discriminate function if this variable is set to zero. Each axis of the function is defined as:

$$XAxis: d = 0.01x; \frac{\partial z}{\partial x} = \frac{z(x+d, y,...) - z(x-d, y,...)}{2d} \quad (2)$$

$$YAxis: \Delta = \frac{z(x, y,...) - z(x=0, y,...)}{z(x, y,...)}$$

where 2·d is the step of the numerical derivative in axis X; x,y,… are attributes of the dataset and z is the discriminate function. On the Y axis, the value of the attribute less itself set to null is used to evaluate its effects in the total value of the discriminate function.

The X axis shows the behavior of the patient, if he is better (negative values) or worse (positive values). The Y axis shows the contribution of the variable to the

improvement of the patient condition (negative value) or to aggravate their condition (positive values). The ideal conditions are both negative values, and the sickly conditions are both positive values.

We termed this graphic as *Disease Pathway Graphic - DPG*, because it reproduces the pathway the patients follow during their recovery in the plane defined by the transformation in Eq. 2.

Data mining effort in atherosclerosis dataset [9].

The use of the attribute in algebraic equations demands a continuous and coherent definition of its values. Continuous variables like cholesterol in mg% present a clear meaning: bigger the module, bigger the concentration of cholesterol.

However, discrete values like married, divorced, single, widower must be sorted in a sequence that shows the stress or constraint like married, divorced, widower, and single. The responsibility, stability, or stress in some criteria should be monotone increasing/decreasing values.

Values not stated were set to zero, because in this case it does not changes the discriminate function value.

The final dataset was obtained combining the record from "Entry" file as the first sequence of measures for each patient, and the posterior records from "Control" executes some change in the last value of the patient record, generating a new record in the database. The final dataset contains 11989 records, each one representing a different sample.

To avoid bias, all attributes were normalized between 0.01 and 1.0 and genetic programming uses this database to obtain the discriminate function.

Experimental results

The discriminate function was able to model 8426 records correctly (71%) and 3563 records wrong (29%). The model obtained is:

(+ (+ (+ (+ (+ (- (- (- alcohol vzdelani) (- (* (* (+ moc chlst) (+ kysmoc (+ (+ (* (- dusnost pivo12) (- alcohol kysmoc)) (- syst1 (- hypll HTD))) (* ldl glykemie)))) (+ (* -3.33355 (* glykemie HT)) (+ (+ (+ (+ imtrv (* -3.33355 (* glykemie HT))) (+ (* glykemie HT) (+ (+ (- hypll HTD) (* ldl glykemie)) (* ldl glykemie)))) (- alcohol vzdelani)) (+ (* ldl glykemie) glykemie)))) (+ (+ (- ICT vinomn) (+ (+ (- (* -3.33355 byvkurak) HT) (* -3.33355 (* glykemie HT))) hypll)) (* (- vyska HTD) (+ dusnost alcohol))))) HT) (* -3.33355 (* glykemie HT))) dobakour) (+ (* (+ imtrv (* -3.33355 (* glykemie HT))) byvkurak) syst2)) (+ (+ (+ vzdelani (+ (* vinomn byvkurak) smoking)) (* (- dusnost pivo12) (- dusnost pivo12))) (+ (+ (+ glykemie (* glykemie HT)) (- hypll HTD)) (* ldl glykemie))))

See Table III for description of mnemonics in the model.

Analysis of accuracy. The result obtained (71%) shows a lower accuracy than the one obtained in breast cancer (96%). To understand its meaning, let us define an average "normal" patient obtained from the dataset of "normal studied group" patients: body mass index (23.28), blood pressure (syst 129 diast 82), cholesterol (216), and triglycerides (147).

Consider the patient 10001 from the risk group. The discriminate function fail to obtain his risk condition (see *** in column OK/NOK Table I) when he has the physical and biochemical parameters stable.

Fig 3 shows the evolution of the discriminate function and its tendencies. It is evident the discriminate function changes its value depending on the patient clinical condition. The doctor should take some action when measured the sample 13 (an increasing tendency since sample 9) to prevent the increase in samples 14, 15, etc.

This is a very important point, because the analysis of discriminate function temporal behavior can forecast future problems of the patient.

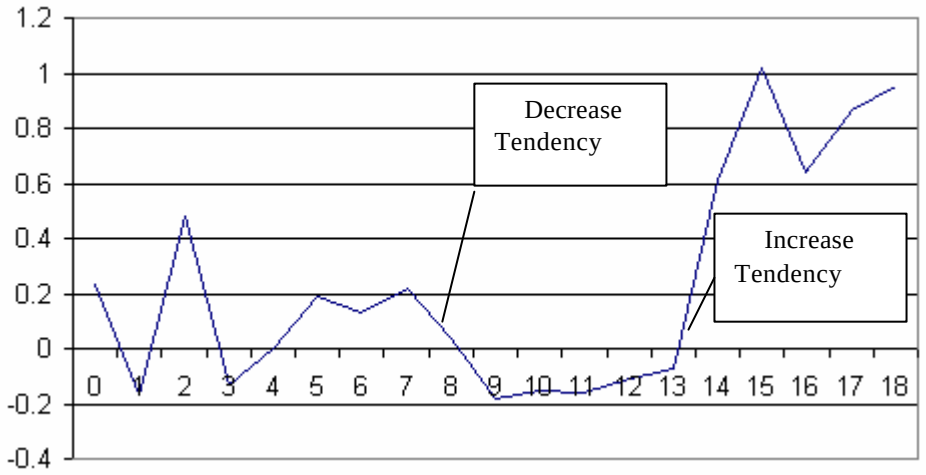


Fig. 3. Evolution of discriminate function over the analysis. The axis represents the discriminate function by sample number (# in Table I).

To study the discrepancies, let us define an index of risk. Each of Table I attributes has an average value for normal patients and let us define a discrepancy threshold if its value is bigger than 10% of the average value. The index is the squared sum of the all discrepancies bigger than 10%.

Table I shows that the discriminate function fails to show risk conditions when the discrepancy is close or bellow the threshold. This means that the discriminate function is not affected by the indistinct classification of data as risk when it is not.

Applying this study for all records where the discriminate function missed (total 3563 cases) is shown in Table II. The discriminate function classify a risk patient with discrepant attributes when the patient is normal and the discrepancy is greater than 0 (total 451 cases), or in the other cases when discrepancy is less than 0 (total 2443 cases).

Table 1. Analysis of physical and biochemical parameters for patient 10001

#	VYSKA	VAHA	SYST1	DIAST1	CHLST	TRIGL	OK NOK	Index
0	169	71	120	85	209	86		0.00
1	169	71	130	90	217	108	***	0.00
2	169	72	**140**	90	232	**389**		1.65
3	169	71	130	90	241	134	***	0.12
4	169	74	**150**	**100**	201	126		0.30
5	169	73	**165**	**105**	235	88		0.40
6	169	71	130	92	216	76		0.12
7	169	72	**175**	**105**	205	74		0.45
8	169	74	140	95	**286**	71		0.38
9	169	75	130	90	189	86	***	0.13
10	169	74	130	90	216	70	***	0.11
11	169	68	140	90	170	89	***	0.00
12	169	71	140	90	193	75	***	0.00
13	169	72	140	90	215	83	***	0.00
14	169	71	120	85	219	87		0.00
15	169	70	**160**	**100**	232	60		0.33
16	169	71	140	90	178	89		0.00
17	169	72	**150**	85	215	84		0.16
18	169	72	**180**	**100**	170	69		0.45

Table II. Analysis of discrepancy.

Patient type	**Discrepancy >0**	**Discrepancy<0**	**TOTAL**
Normal	**451**	102	553
Risk	426	**2423**	2879
Pathologic	107	**17**	125
Death	3	**3**	6

Hence, the method found correct values in 8426 cases when running GP to obtain the discriminate function, and is correct in 451+2443 cases where the discriminate value do not gave the classification due the condition of the patient do not agree with his classification. The total of correct values (8426+451+2443=11320) represents 94% of the total number of records (11989).

The selection of all individuals where the discriminate transited between positive/negative values are 202 and 269 for decrease and increase tendency. The study of the behavioral attributes to establish the causes of this change is available in Table III.

First of all, the transition probability in both directions is the same. This is an important point of discussion for the workshop.

There are several behaviors that can help in the transition: be married, high education, partly independent, mainly sits in the job (???), moderate physical activity

after job, drink only occasionally, do not smoke, sugar not affect, less than 2 glasses of coffee a day, and less than 2 glasses of tea a day.

Table III Attributes influence in discriminate transition

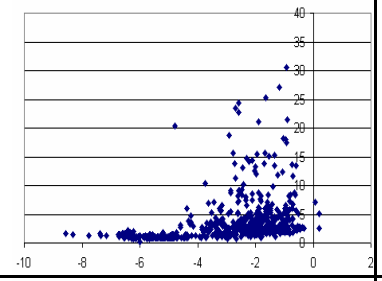
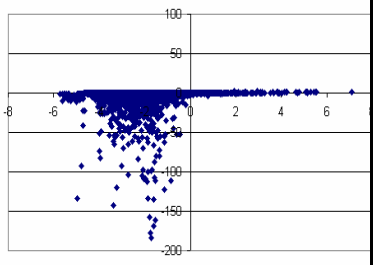
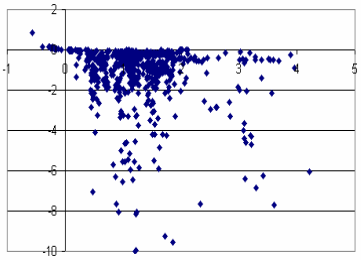
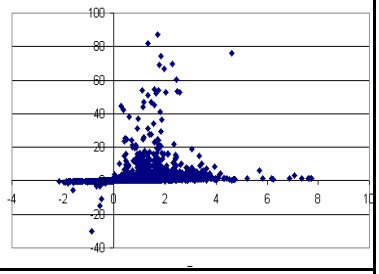
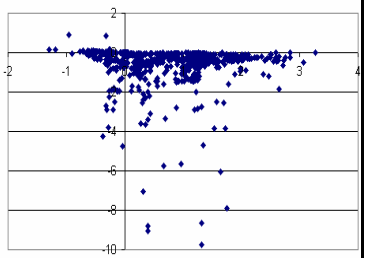
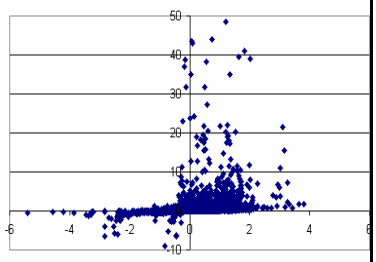
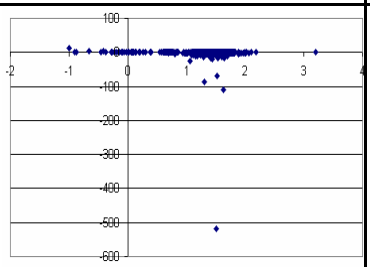
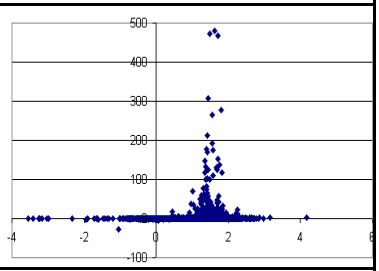
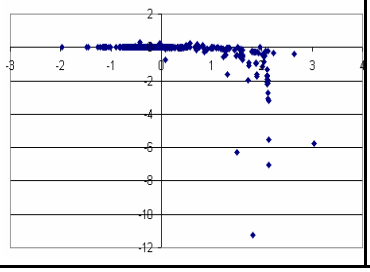
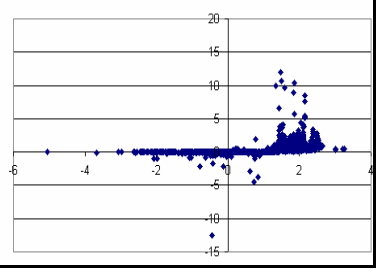
Attribute		%Dec	%Incr	Attribute		%Decr	%Incr
STAV-married status	Married	87	89	VZDELANI- reached education	basic	0	0
	Divorced	1	2		apprentice	15	16
	Widower	1	1		Secondary	38	35
	Single	9	7		University	46	48
ZODPOV-responsibility in a job	Managerial	21	19	TELAKTZA-physical activity in a job	mainly sits	70	67
	partly independent	39	31		mainly stands	10	13
	Others	27	24		mainly walks	15	15
	Pensioner	11	24		heavy loads	3	3
AKTPOZAM-physical activity after a job	mainly sits	18	17	ALKOHOL - drinking of alcohol	No	10	10
	Moderate	68	73		Occasionally	69	66
	Great	12	9		Regularly	20	23
KOURENI - intensity of smoking	No	64	65	CUKR - daily consumption of sugar lumps	No	20	21
	1-4	5	4		1-3	23	24
	5-14	10	6		4-6	31	34
	15-20	14	11		7-9	12	12
	>21	5	10				
	Cigars/pipes	0.5	2				
KAVA - daily consumption of coffee	No	27	32	CAJ - daily consumption of tea	No	40	36
	1-2	51	46		1-2	52	55
	>3	22	21		>3	7	8

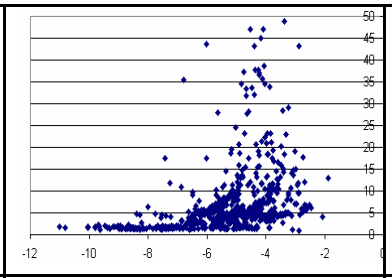
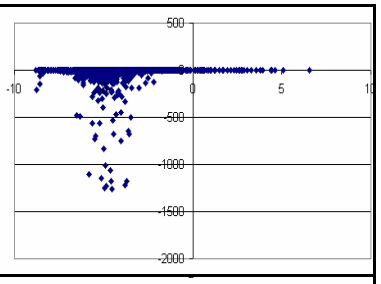
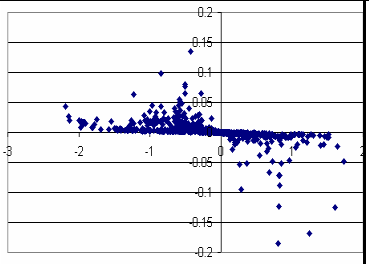
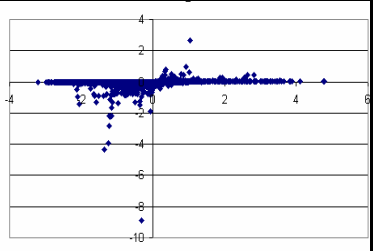
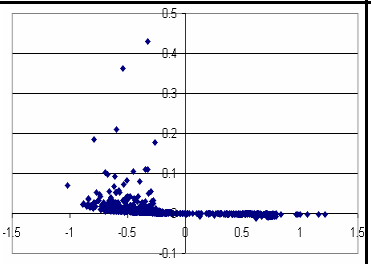
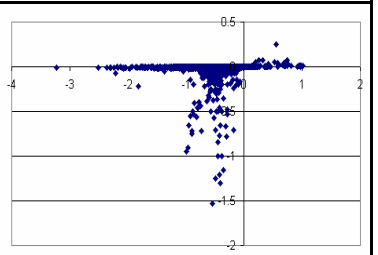
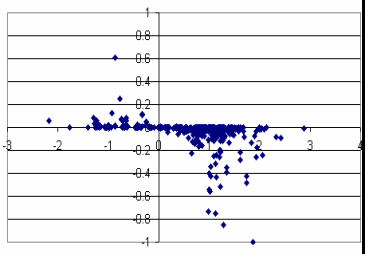
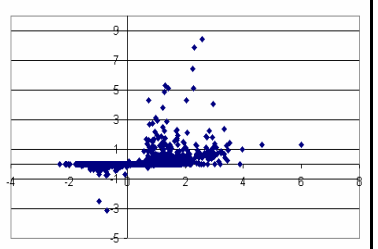
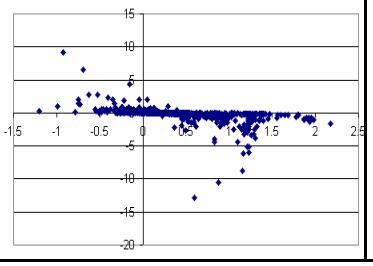
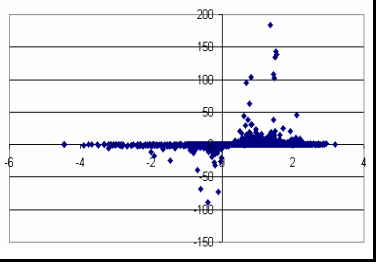
Analysis of the influence of each attribute in the discriminate function.

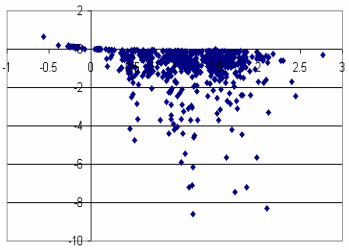
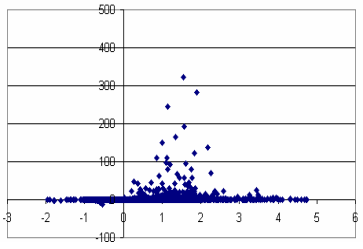
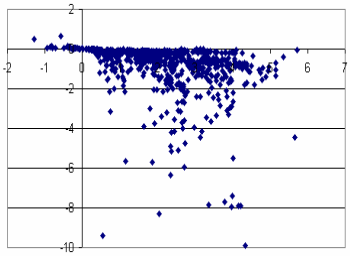
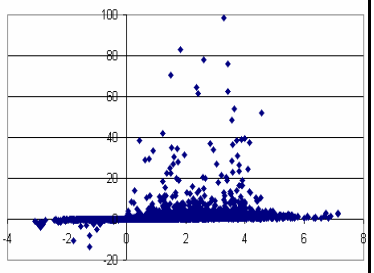
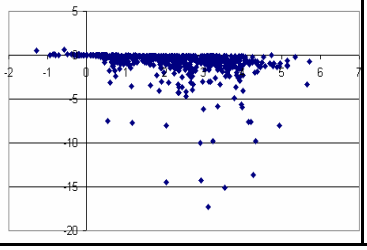
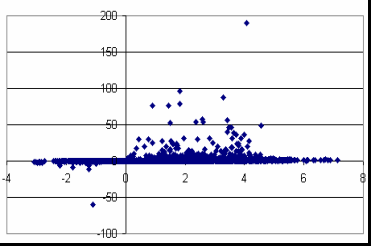
Table IV shows the DPG of each relevant attribute in the discriminate function value for the correct predictions of the GP run.

The analysis of the Table IV shows that discriminate function cluster the data, in two different groups (risk and not risk). The supervision of the DPG patient pathway would contribute to forecast the evolution of patient conditions.

Table IV DPG for main attributes of discriminate function

Attribute	Health Patient	Risk Patient
VZDELANI - reached education		
ZODPOV - responsibility in a job		
TELAKTZA - physical activity in a job		
ALKOHOL - drinking of alcohol		
PIVO12 - beer 12°		

HT - hypertension		
HTD - diet in HT		
IMTRV – before how many years myocardial infarction has appeared		
DUSNOST - asthma		
VAHA - weight (kg)		

SYST1 - blood pressure I systolic (mm Hg)		
CHLST - cholesterol in mg%		
MOC - urine		

Summary and conclusions

This essay studies how to obtain the mathematical model of the disease, and classify the patient condition. The method is able to deal with the problem of patient label (risk even when the condition is not risk).

The accuracy of the method is around 94%, if considered the 10% average deviation indexto classify the patient condition.

The forecast of the patient condition can be done by the discriminate function monitoring and analysis of its tendency. This is a procedure easy to be executed in any laboratory.

Topics for discussion during the workshop.

The discriminate function is able to classify patients with different level of risk with good accuracy. We would appreciate very much discuss what information is able to be obtained from the DPG and how to analyze it, based in the information obtained by others data mining techniques.

Other important question is why the decrease and increase tendency presents the same number of events for the same conditions? I guess that there are different causes for increase and decrease.

Discriminate function is a powerful tool, and the interpretation of its results would contribute in the understanding of many diseases where the method was applied.

References

1. Werner,J.C.; Kalganova,T.; "Disease modeling using Evolved Discriminate Function". LNCS 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.
2. Werner,J.C.; Fogarty,T.C.; "Severe diseases diagnostics using Genetic Programming." Intelligent Data Analysis in medicine and pharmacology – IDAMAP2001; September 4th, 2001 London http://magix.fri.uni-lj.si/idamap2001/scientific.asp
3. Werner,J.C.; Fogarty,T.C.; "Genetic programming applied to Collagen disease & thrombosis." in PKDD 2001 Challenge on Thrombosis data – Germany/ Freiburg September 3-7, 2001.
4. HOLLAND,J.H. "Adaptation in natural and artificial systems: na introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992.
5. GOLDBERG,D.E. "Genetic Algorithms in Search, Optimisation, and Machine Learning." Reading, Mass.: Addison-Whesley, 1989.
6. CHAMBERS,L.; "The practical handbook of Genetic Algorithms" Chapman & Hall/CRC,2000.
7. KOZA,J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge,Mass.: MIT Press, 1992.
8. LilGP "Genetic Algorithms Research and Applications Group (GARAGe)", Michigan State University; http://garage.cps.msu.edu/software/lil-gp/lilgp-index.html
9. Discovery Challenge 2003, ECML/PKDD 2003 Conference, Dubrovnik, Croatia, 22-26 September 2003.