Werner J. C. and T. Kalganova (2003) Disease modelling using Evolved Discriminate Function. Proc. of the 6th. European Conference on Genetic Programming, EuroGP2003, Essex, UK. Published by Springer-Verlag. Vol. 2610. pp. 465-473.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

Disease modeling using Evolved 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. Precocious diagnosis increases the survival time and patient quality of life. It is a binary classification, exhaustively studied in the literature. This paper innovates proposing the application of genetic programming to obtain a discriminate function. This function contains the disease dynamics used to classify the patients with as little false negative diagnosis as possible. If its value is greater than zero then it means that the patient is ill, otherwise healthy. A graphical representation is proposed to show the influence of each dataset attribute in the discriminate function. The experiment deals with Breast Cancer and Thrombosis & Collagen diseases diagnosis. The main conclusion is that the discriminate function is able to classify the patient using numerical clinical data, and the graphical representation displays patterns that allow understanding of the model.

Introduction

Britain has the worst survival rates for cancer of any nation in the western world. The five-year survival rate for breast cancer, if diagnosed early, is 78% in the UK, compared with 97% in America and 93% in the rest of the Europe [1]. The study found many British patients were diagnosed only when their cancer was at an advanced stage and that was more difficult to treat. Early diagnosis increases the survival chances.

One way to improve early diagnosis is to develop modeling techniques able to identify imperceptible patterns from datasets and support decision making. In practice, what has to be done is to input the patient records into the model and obtain a forecast of the diagnosis.

Extensive work has already been carried out in this area (see Table 1). The adopted paradigms adapt parameters and threshold values in a pre defined fixed mathematical structure.

Analysis of existing approaches (Table 1) show two main drawbacks: (1) the created model of the knowledge does not take into account false negative events. (2) the adaptation of a previous defined model structure (such as If-then-else, or Neural Network with back propagation, etc) to the problem. The first drawback is that false negative and false positives have the same weight in the obtainment of the model.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

While false positive is a safe condition to the patient because new clinical analysis will be carried on, false negative is a dangerous postponing of the diagnosis and decreases survival chances. The consequence of a second bottleneck is that some problems will not have the best model, e.g. they will have an approximate solution using the structure. This approximation masks the dynamics of the disease.

Table 1. Knowledge representation for medical diagnosis (KR – knowledge representation, NN – Neural Network, SVM - Support Vector Machines, EA – Evolutionary algorithm, DT – Decision Trees, FZ – Fuzzy, GP – Genetic Programming).

Author	Algorithm	Description
Kononenko 2001 [2]	naïve Bayesian classifier, NN and DT	*KR: statistical parameters, parameter adaptation, and rules* Only decision tree builders are able to select the appropriate characteristics (performance, transparency, explanation, reduction, and missing data handling).
West 2000 [3]	NN	*KR: parameter adaptation* Several different neural networks were applied and will be used as reference to compare with our approach
Setiono 1996 [4], 2000 [5]	rules from NN	*KR: rules to represent the parametric model* Extraction of rules from a trained NN to overcome its black box concept
Flach 2001 [6] Joachins[7]	SVM	*KR: Geometrical approach*. Optimum margin classifier + kernel. Training examples are linearly separable and try to obtain a hyper plane with maximum margin from positive and negative points
Land Jr 2002 [8]	EA to configure SVM	*KR: optimal hyper planes geometry* Improvement of the specificity by 45.3% at 100% (missing no cancer) when compared with iterative method
Pendharkar 1999 [9]	machine learning	*KR: rules* Rules and the observation of patterns and knowledge acquisition for various knowledge base systems. The application in breast cancer diagnosis shows that the method is a viable tool
Nauck 1999 [10]	FZ	*KR: linguistic rules* Fuzzy rule based classifier with simple linguistically interpretable rules
Freitas 2002 [11]	EA	*KR: rules coded in the chromosome.* Chromosome codes rules If-Then-Else with the attributes and the classification can be understood
Pena-Reyes 1999 [12]	EA + FZ	*KR: rules with optimal transition values*. Breast cancer diagnosis with high performance
Proposed method	EA(GP)	*KR: mathematical model of the disease*. Obtain the discriminate function for the disease to classify the patients.

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 that should be used in the model assembly, which results in an enormous degree of freedom. 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 paper the origin was adopted as a threshold: positive values mean an ill patient and negative values a healthy patient.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

A multiplicative weight (termed punishment) is introduced to give more priority to false negatives. It guarantees minimal false negatives, which costs accuracy in true negative values. Again, this is a safe condition for the patient.

The experimental results prove the reliability of the proposed approach. However, more than 95% accuracy is not enough if the user is not able to understand how the algorithm works and what they have learned. To overcome this difficulty, a new graphical representation is proposed to analyze the discriminate function and show the contribution of each attribute in the fitness function.

The experimental results used two datasets to evaluate the method: one is the Wisconsin Breast Cancer dataset and the other is the Collagen Disease and Thrombosis dataset.

Genetic Programming

GP is an optimization algorithm which mimics the evolution and improvement of life through reproduction. 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 [13], [14], [15], [16]. Specialized Markov Chains underline the theoretical bases of this algorithm, changes of states and searching procedures.

The software we have developed is an adaptation of LilGP [17], where GP is structured in a pre-compiled library, with other artificial intelligence procedures, such as NN, FZ, adaptive algorithms, etc. Outputs are written in Excel XLS format direct from the program, to generate an accessible and functional Human-Computer Interface (HCI).

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($\cdot$), 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.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

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 divided in two parts: one is for training and the second is for validation. The training dataset is used to obtain the model and the validation dataset is used to measure the accuracy of the model with data that was not used in training.

The fitness function evaluates how good the diagnostic model coded in chromosome is, over all training dataset using Receiver Operating Characteristics (ROC) [18].

ROC criterion value is sliding in the output projection and the number of true negative (N_{TN}), true positive (N_{TP}), false negative (N_{FN}), and false positive (N_{FP}):

$$\alpha = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad \beta = \frac{N_{TN}}{N_{TN} + N_{FP}} \tag{1}$$

where á is the *Sensitivity*, and â is the *Specificity*. Sensitivity is the probability that a test result will be positive when the disease is present (true positive rate, expressed as a percentage). *Specificity* is the probability that a test result will be negative when the disease is not present (true negative rate, expressed as a percentage).

The fitness function *F* used in the disease diagnostic is the accuracy of the model, with a weight over false negatives predictions:

$$F = \frac{N_{ok}}{N_{ok} + N_{FP} + \sigma * N_{FN}} \tag{2}$$

where ó is the overprice for false negative (high risk condition), or punishment weight, N_{ok} is the number of correct forecast, N_{FP} is the number of false positives and N_{FN} is the number of false negatives.

Diagnosis of severe diseases using discriminate function

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 graph 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:

$$XAxis: \delta = 0.01x; \frac{\partial z}{\partial x} = \frac{z(x+\delta, y, ...) - z(x-\delta, y, ...)}{2\delta} \tag{3}$$

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

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

where 2· ä 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. 3.

Experimental results

The following subsections present the experimental results for breast cancer from Wisconsin University [19], [20] and Collagen Disease and Thrombosis from Chiba Hospital [21], [22]. In both cases, GP was applied to obtain the discriminate function with the training dataset and the test is done applying a validation dataset to evaluate its effectiveness.

Breast-cancer testing is an important application because it is crucial to develop a reliable but inexpensive test to identify women with high risk for a more expensive and accurate clinical procedure.

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.

Experiment 1: Breast cancer diagnostic. The Wisconsin Diagnostic Breast Cancer [19], [20] contains 679 events (236 ill and 443 healthy records) without any missing values. The dataset contains the following attributes: Clump Thickness (clth), Uniformity of Cell Size(uncz), Uniformity of Cell Shape (uncs), Marginal Adhesion (mara), Single Epithelial Cell Size (sepc), Bare Nuclei (barn), Bland Chromatin (blac), Normal Nucleoli (norn), Mitoses (mito), Class (2 benign 4 malignant). Each attribute is an integer between 1 and 10.

An input routine reads the data and stores it in memory and fitness function evaluates the accuracy of the discriminate function of each individual. The discriminate function is checked against the type of tumor (benign or malign) to find the fitness function (Eq. 2).

The first study was the effect of punishment factor in the sensitivity. The complete dataset is modeled using GP to obtain the discriminate function with different values of punishment in the fitness function (Eq. 2). The results shown in Table 2 present the effect of different values of punishment weight in sensitivity and specificity.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

The highest value of sensitivity (with lower false negative) is obtained when punishment is equal to 10 (see bold values in Table 1). The occurrence of false negatives decays for values greater than or equal to 5, and oscillates around 4 false negative values, without vanishing for 500 generations of 100 individuals.

Table 2. Study of punishment value (column Punishment) for breast cancer modelling. Bold value point to the highest sensitivity value (**á** is the sensitivity and **â** is the specificity).

Punishment	N_{TN}	N_{FP}	N_{TP}	N_{FN}	**á**	**â**
1	425	18	226	10	95.7	95.9
3	417	26	227	9	96.1	94.1
5	412	31	231	5	97.8	93.0
10	**420**	**23**	**234**	**2**	**99.1**	**94.8**
15	402	41	231	5	97.8	90.7
20	406	37	233	3	98.7	91.6

Punishment equal to 10 is used in a run to obtain the model where the number of false negatives is null. This experiment will be used to generate the Disease Pathway Graphic. The maximum number of generations was 2000, and the best solution was obtained after 543 generations. The discriminate function was:

```
(barn + uncs) (-34.72 + barn + clth + barn clth - barn sepc + norn / uncs + clth * clth  /
((85.53/ blac - 2 blac + mara - mito + blac sepc) (sepc + uncs)) + blac (sepc + uncs) +
                barn uncz + (mara - uncs)/ (sepc/norn + uncz))
```
(4)

The number of true negatives is 426, false positives is 17, and true positives is 236. The accuracy is 97.5%.

Breast cancer was studied by many authors and is a benchmark. We will follow the same methodology of West [3] to compare the results and accuracy. To obtain the discriminate function for diagnostics proposes 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 original databases used in the test stage.

Table 3. Different approaches to cancer diagnosis [3]

Method	**OK (%)**	**% False negative**	**% False positive**
Multilayer perceptron	0.957206	0.087448	0.018594
General regression	0.967647	0.054393	0.020408
Radial basis function	0.970441	0.030126	0.029252
Mixture of experts	0.962941	0.062762	0.023129
Logistic regression	0.9633968	0.0711297	0.018018
Logistic	0.972182	0.029289	0.027027
K search neighbor	0.967789	0.033473	0.031532
Kernel	0.95022	0.117155	0.013514
GP (proposed)	**0.963235**	**0.008368**	**0.05180**

The software should run until a model is found without false negative. However, it is not a guarantee that new data will be modeled without false negatives. In this experiment we used the same parameters of the study of punishment weight (100

individuals, 500 generations, 60% crossover probability and 20% mutation probability). This will give an idea of the usual level of false negatives.

The comparison with several techniques from [3] shows that all these techniques have more false negatives than the discriminate function using GP. Table 3 shows the different results for different data mining techniques.

The occurrence of false negative is the least value of available approaches without compromise of the total accuracy, paying the price of a greater false positive than the other approaches. For example, the average false negative of the other methods is 0.60%. If the method were applied in London (9 million people) there were 543,000 patients false negative against "only" 74,700 using our method. Let us consider that the algorithm can be improved for a null false negative model.

This experiment shows a good level of accuracy with low false negatives. The algorithm can model the disease with an algebraic equation which reproduces the dynamics of the disease. However, the model (Eq. 4) does not allow the user to understand the importance of each attribute in the diagnostic, and the effect it causes in the model.

Analysis of the disease model. To study the disease dynamics model of breast cancer, we use all true negative events and all true positive events to draw the graphics of disease pathway (transformed with Eq. 3) in Fig. 3 obtained with Eq. 4. There is a different pattern for ill and health patients.

Healthy patients are clustered close to the origin, while ill patients are spread under a pattern over the first quadrant. The end of the scale is fixed to the same value to all variables to show the comparative behavior.

The attributes "Marginal Adhesion (mara)" and "Normal Nucleoli (norn)" are distributed around the origin and do not influence the diagnosis at all. The dataset fail to provide a history of each patient, to plot its temporal evolution. However, this can be analyzed using the thrombosis dataset, if the missing values problem were solved. In this case all 57,545 records would be used in the experiment, and not only 261 or 1988 records.

Experiment 2: Collagen disease and Thrombosis diagnostic. The purpose of this experiment is to apply the method to a more complex dataset [21], [22]. There are three degrees of disease diagnostic: mild, severe and most severe. The dataset differs from breast cancer because it contains missing values, few elements for positive diagnosis, more than one degree of disease, and noisy data.

Our approach has been applied to the dataset available with information on concentration compounds in the blood exam (lab – 57,545 records - and antibody – 773 records - exams). To train the software, the same examination date and patient was selected from both datasets forming a dataset with 261 records (231 none and 30 yes).

GP used 11,072 generations to find the best solution (the limit was 40,000 generations), with a population of 100 individual, 60% crossing over probability, and 20% mutation probability. In this experiment the punishment weight for false negative is 20.

To test the discriminate function (validation) with data that was not used in training, we accept records where Lab and antibody date exams differs into one month totalizing 1988 records (1564 none and 424 yes).

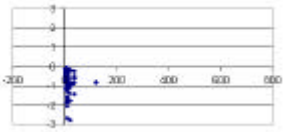
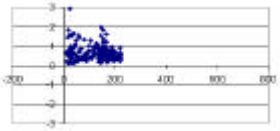
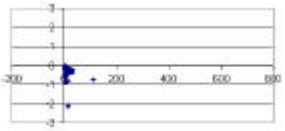
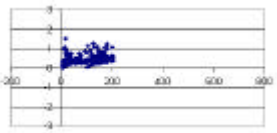
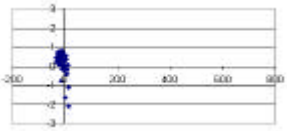
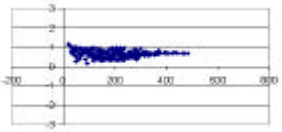
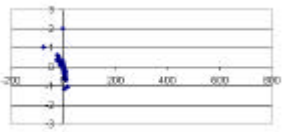
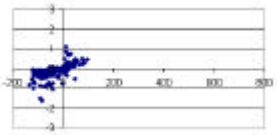
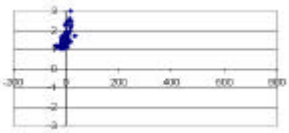
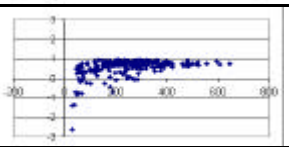
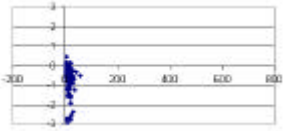
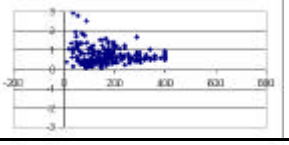
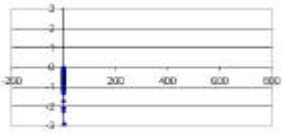
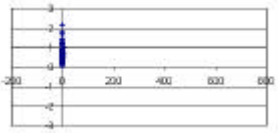
Attribute	Negative (Health)	Positive (ill)
Clump Thickness (clth)		
Uniformity of Cell Size (uncz)		
Uniformity of Cell Shape (uncs)		
Single Epithelial Cell Size (sepc)		
Bare Nuclei (barn)		
Bland Chromatin (blac)		
Mitoses (mito)		

Fig. 1. Different behavior of the variables in the disease pathway graphics.

This dataset contains missing values and undefined values such as < 3.0. Missing values were filled with the average of each missing attribute. Undefined values were replaced by the threshold value.

Table 4 shows that there is consistency between the training and validation sets, but it is possible to see that the “none” events pay the price of the punishment into the false negative case.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

Table 4. Discriminate function for Collagen Disease.Total number of events (N^{Tot}), number of correct predictions (N^{CP}) and percentage of correct prediction over total number (% N^{CP}) for each disease degree in training and validation datasets runs. GP parameters are population size (λ), number of generations (N_{gen}), crossover probability (p_c), and mutation probability (p_m).

GP parameters			Training			Validation		
		Diagnostic	N^{Tot}	N^{CP}	% N^{CP}	N^{Tot}	N^{CP}	% N^{CP}
λ	100	None	231	172	**74**	1564	968	**61**
N_{gen}	3924	Mild	1	1	**100**	1	1	**100**
p_c	60 %	Severe	11	10	**90**	250	231	**92**
p_m	20 %	Most severe	18	18	**100**	173	168	**97**

In validation results, the number of false negative is 24 (1.2%), with represent risk for the patient. The accuracy of the method is 76% for training and 68% for validation.

There are two possible explanations for these results: the effect of missing values and the use of average values; the low number of diseases in the training dataset (only 11% of the cases are ill patients) and the low number in the training dataset (261 records).

However, the method was able to obtain a model for the disease with low false negative.

Summary and Conclusion.

This paper presents an approach for classification using a mathematical discriminate function. To reduce false negative, different punishment values were tested. It shows that its value is critical below a threshold and does not affect the result accuracy after this point.

With the punishment value, we obtained the discriminate function for breast cancer and collagen disease with good accuracy, showing that the method can be applied to model diseases using an algebraic equation of the attributes. To extract information from the model, we proposed a graphical representation of the discriminate function that allows visualization of each attribute and its effects in the discriminate function. The graphical presentation of each variable gave a better understanding of each attribute contribution and would help to clarify the knowledge acquire by the model. Due to the capability to predict the disease, the model contains the dynamics of the disease under study and this approach can contribute to the improvement of diseases treatment. Thanks to Susan McCracken and Owen Parry for proof reading in this paper.

References

1. Newman,M.; "UK's cancer death rate is worst in the world"; Metro News, Tuesday, July 2, 2002.
2. Kononenko,I.; "Machine learning for medical diagnosis: history, state of the art and perspective"; Artificial Intelligence in medicine 23:89-109, 2001.

Lectures notes in Computer Science 2610, Proceedings 6th European Conference, EuroGP 2003, Essex, UK, April 14-16, 2003.

3. 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.
4. Setiono,R.; "Extracting rules from pruned neural networks for breast cancer diagnosis" Artificial Intelligence in Medicine 8(1):37-51, 1996.
5. Setiono,R.; "Generating concise and accurate classification rules for breast cancer diagnosis"; Artificial Intelligence in medicine 18:205-219, 2000
6. Flach,P.A.; "On the state of art in machine learning: a personal review"; Artificial Intelligence 131:199-222, 2001.
7. Joachins, T.; "Tutorial Support Vector Machines" In Internet http://www.afia.polytechnique.fr/CAFE/ECML01/SVM.html
8. Land Jr., W.H.; Lo,J.Y.; Velazquez,R.; "Using evolutionary programming to configure support vector machine for the diagnosis of breast cancer". In Dagli,C.H. et al (Eds) Intelligent engineering systems through artificial neural networks ANNIE'2002, Volume 12, Smart engineering system design, ASME Press, New York, 2002.
9. Pendharkar,P.C.; et al; "Association, statistical, mathematical and neural approaches for mining breast cancer patterns"; Expert Systems with Applications 17:223-232, 1999.
10. Nauck,D.; Kruse,R.; "Obtaining interpretable fuzzy classification rules from medical data"; Artificial intelligence in medicine 16:149-169, 1999
11. Freitas,A.A.; "Data mining and knowledge discovery with Evolutionary Algorithms"; Springer 2002.
12. Pena-Reyes,C.A.; Sipper,M.; "A fuzzy-genetic approach to breast cancer diagnosis"; Artificial intelligence in medicine 17:131-155, 1999.
13. HOLLAND,J.H. "Adaptation in natural and artificial systems: na introductory analysis with applications to biology, control and artificial intelligence." Cambridge: Cambridge press 1992.
14. GOLDBERG,D.E. "Genetic Algorithms in Search, Optimisation, and Machine Learning." Reading, Mass.: Addison-Whesley, 1989.
15. CHAMBERS,L.; "The practical handbook of Genetic Algorithms" Chapman & Hall/CRC,2000.
16. KOZA,J.R. "Genetic programming: On the programming of computers by means of natural selection." Cambridge,Mass.: MIT Press, 1992.
17. LilGP "Genetic Algorithms Research and Applications Group (GARAGe)", Michigan State University; http://garage.cps.msu.edu/software/lil-gp/lilgp-index.html
18. Bradley, A.P.; "The use of the area under the ROC curve in the evaluation of machine learning algorithms"; Pattern Recognition, 30(7):1145-1159, 1997.
19. WDBC Dr. William H. Wolberg, General Surgery Dept.,; W. Nick Street, Computer Sciences Dept.; Olvi L. Mangasarian, Computer Sciences Dept.; University of Wisconsin http://www.ics.uci.edu/~mlearn/MLRepository.html
20. 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
21. 5th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD'01) Challenge on Thrombosis data – Germany/ Freiburg September 3-7, 2001
22. 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.