A new medical test for atherosclerosis detection GeNo (PKDD 2003 Challenge)

Jérôme Azé, Noël Lucas, and Michèle Sebag

LRI
CNRS UMR 8623
Université d'Orsay
F-91405 Orsay
{aze, lucas, sebag}@lri.fr

Abstract. This paper is concerned with the analysis of atherosclerosis factors. We present a new medical test, which is very cheap and allow the pratician to take better decision when treating a patient for atheroscleoris. We used genetic algorithm to compute the better test given a specific fitness fonction. This fonction is based on the Area Under the ROC curve. The test have been validated using the initial examination and the long term observation.

1 Introduction

Atherosclerosis : name of the process in which deposits of fatty substances, cholesterol, cellular waste products, calcium and other substances build up in the inner lining of an artery. This buildup is called plaque. It usually affects large arteries. Atherosclerosis is a slow, complex disease that starts in childhood and insidiously progresses when people grow older.

Plaques can grow large enough to significantly reduce the blood's flow through an artery. But most of the damage occurs when they become fragile and rupture. Plaques that rupture cause blood clots to form that can block blood flow or break off and travel to another part of the body. If either happens and blocks a blood vessel that feeds the heart, it causes a heart attack. If it blocks a blood vessel that feeds the brain, it causes a stroke. And if blood supply to the arms or legs is reduced, it can cause difficulty walking and eventually gangrene.

Unfortunately, to date, physicians do not have treatment for atherosclerosis. As prevention, they only propose reducing the well known controllable risk factors for causing damage to the arterial wall include elevated levels of cholesterol and triglyceride, high blood pressure, tobacco or diabetes. More, there is interactivity beetween these factors and nobody knows what exactly happens. So, there is no true primary prevention.

Studies help physicians in discovering factors and understand how to deal with them : what to do and when. Four databases related to the risk factors of atherosclerosis have been made publicly available for the PKDD 2003 Challenge.

Our work proposes in first answers to a part of the analytic questions related to the *Entry* database examination, *in fine*, gives to the physician a simple, effective and predictive test to know if a man has a high risk to have cardio-vascular disease from his atherosclerosis.

2 Data Preparation

It is worth noticing that one difficulty related to the data sparsity has been addressed by condensing sets of attributes into a few global boolean attributes. Actually, due to the great level of detail of the medical campaign, the initial database involved attributes which were hardly significant from a statistical perspective (e.g. the diseases of the patient's fourth sister were hardly informed 4 times over 1419 patients), and they were therefore discarded [10]. To resist this sparsity phenomenon, in our previous contribution to the Atherosclerosis PKDD Challenge [7], we forged specific attributes inspired from ANAES recommendations[1]; and we were happy to see that the new organization of the data followed this line in 2003.

Sometimes, attributes of the *Entry* database present missing values. These missing values have been replaced by the mean value of the attribute. So we have a complete database without any missing value.

Not any more work have been done for this database except for the toxicological and physical attributes.

Toxicological problems

Two real-valued variables have been considered, mesuring the volume of alcohol ingested and the number of cigarettes smoked.

Regarding alcohol, three factors have been taken into account: the equivalent amount of alcohol (expressed in g/l); the nature of alcohol (wine being considered less harmful for cardio-vascular diseases than beer, the equivalent alcohol amount has been divided by two); and the patient's weight, as normalizing factor.

Regarding tobacco, two factors have been taken into account: the amount of cigarettes smoken (considering that smoking cigars or pipe is equivalent to 1/2 packet per day) and the number of years. The presence of an interruption has been modelled through a multiplicative factor. The tobacco intoxication factor is multiplied by .8 if the interruption started less than one year ago and by .4 otherwise.

Sport practice and activities

A numerical variable was created to account for the energy, measured in kCal per day, spent daily by each individual during his work, in order to go to work, and during sport activities.

Standard evaluations were used (e.g. walk is account for 170 kCal/hour; moderate activity is 150 kCal /hour; sleep is 60 kCal/hour). Missing information

[1] French National Agency of Accreditation and Evaluation in Heatlh

is accounted for as default energy consumption (90 kCal/hour) or duration (e.g. 1 hour for transportation time).

Each risk factors have been normalized to ensure egality between them.

3 Answers to an analytic question

3.1 General answer

The question we tried to answer is to know if they are any differences between men of the two risk subgroups, those who came down with the observed cardiovasculare diseases and those who stayed healthy.

We want to see if they are differences between these two populations for these factors: tobacco, alcohol, age, bmi, cholesterol, triglycerides, systolic, diastolic, social (marital status and educational level), physical activities.

All the patients who do not came down with one of the observed cardiovascular diseases, are considered as healthy patients. We are interested in finding differences between the two populations, regarding one of the twelve attributes mentionned in the challenge's question.

For each attribute, we sort the patients with respect to their attribute's value. And giving these two populations, we used a Wilcoxon's test [9] to see if the two populations came from the same global population or not.

At a 1% level of significance, the wilcoxon's test show that the two populations are different for the attributes: *age, educational level, physical activities at work, tobacco, cholesterol, systolic, diastolic* and not different for the other attributes.

These observed differences may be used to confirm the fact that patient belonging to the risk group are different from the rest of the population. And that these differences could be observed using the *Entry* database.

3.2 Detailled answer

More interestingly, we tried to see if there are any differences between men of the two risk subgroups themselves. We make the difference between these men (all classified 'risk' in *Entry*) by seeing their future in the databases *Control* and *Death* .

Until the end of the paper, in their future, patients are considered as ill if (HODN1 or HODN2 or HODN3 or HODN11 or HODN11 or HODN12 or HOD13 or HODN14 or HODN21 or HODN23) is (are) positive or if they are in *Death* database (but in this case, from atherosclerosis only (i.e cardiovascular disease)). The healthy patients, belonging to the risk group, are neither ill nor death and are not (HYPERSD or HODN4 or HYPTGL or HYPCHL) positive.

With these definitions, we have, for all risk patients in *Entry* the following distribution (see Table 1).

Then, we want to see if there are any differences between men becoming ill and those staying healthy, and if these differences can be observed using the *Entry* database.

	konskup = 3	konskup = 4
healthy patient	69	65
ill or dead patient	66	74

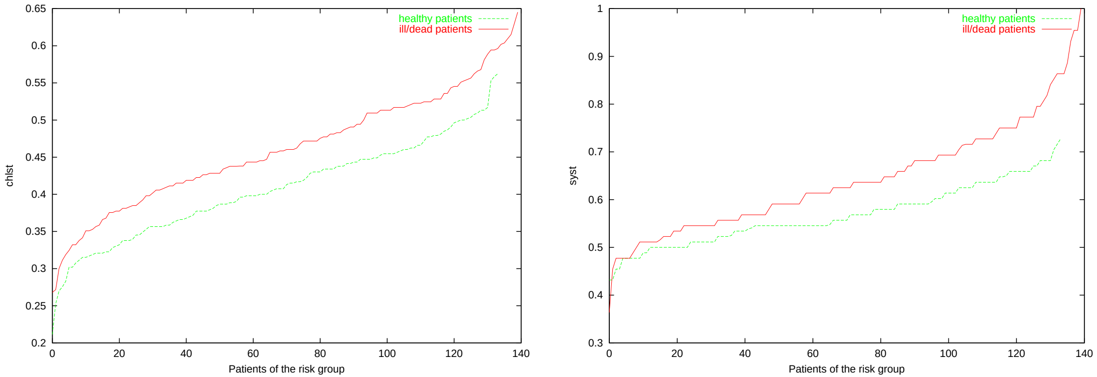
Table 1. Patients from the risk group.

To achieve this goal, we propose to use univaried and bivaried data analysis as described in the next section.

Univaried Analysis

In this analysis, we show for each risk factor its impact on the two risk groups of patients. Each factor has been analysed independently of the other ones.

To plot the curves, patients were sorted by increasing values of the studied factor. Figure 2(a) and 2(b) present impact of cholesterol and systolic blood pressure for healthy and ill patients.



(a) cholesterol (b) systolic blood pressure

Table 2. Patients of the risk group

These curves, for these two factors, show very well that in *Entry* database there are already high differences between the two groups.

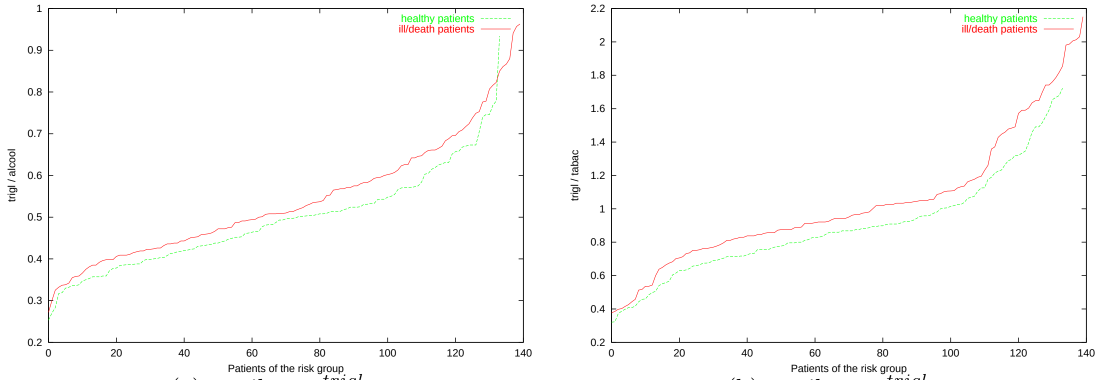
To study the interaction between two risk factors, we use bivaried analysis.

Bivaried Analysis

Based on [6], we combined attributes using multiplication or division operator. For each combination, we obtain a new attribute. Doing this we want to show the complex interaction of the different couples of risk factors.

It is well known in medical science, that alcohol and triglycerides are closely connected. The Figure 3(a) presents this interaction and prove that our model works well.

The second Figure 3(b), presents the interaction between triglycerides and tobacco. Looking at this figure, we can conclude that for two patients having the same tobacco intoxication, the one with the lower triglycerides level have lower risk of atherosclerosis.



(a) attribute $\frac{trigl}{alcool}$ (b) attribute $\frac{trigl}{tobacco}$

Table 3. Patients of the risk group

The conclusion of this work is that first the interaction between the different risk factors are not simple, and secondly it is impossible to explore manually all the interactions of two or more risk factors.

We proposed to use genetic approach to explore all this possible interactions. Using this approach, we have built a medical test which allow us, as present in the next sections, to be more precise for the patient of the initial risk group.

To built the medical test, we used a genetic algorithm which try to find the test having the better ROC Curve.

4 ROC Curve and Learning Stability

This section first describes ROC analysis, then relates this analysis to the requirement of learning stability.

Only binary learning will be considered in the rest of the paper, for the sake of clarity; nevertheless, ROC analysis has been extended to multi-class learning [5].

4.1 ROC Analysis

Let $\mathcal{L}$ be an algorithm, and consider all classifiers obtained from $\mathcal{L}$ by varying its learning parameters on an application domain. This set of classifiers can be visualized as a 2D curve in the FP/TP plane, known as ROC curve associated to $\mathcal{L}$. The ROC curve gives a global overview about the induction tradeoff for the learning algorithm on the application domain; it allows the expert to inspect how the true positive rate must be degraded in order to improve the false positive rate and *vice versa* (Fig. 1). Such a visual inspection is well suited to domains with skewed data distributions, as the false and true positive rates are normalized.

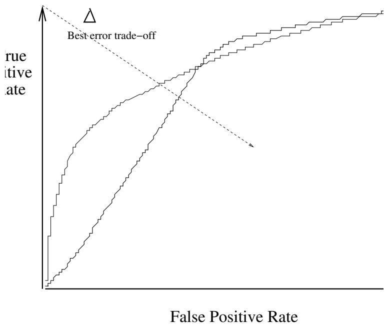


Fig. 1. Comparative evaluation of two algorithms from their ROC curves

Let us interpret the misclassification cost ratio r in the ROC landscape. Clearly the ideal classifier corresponds to point $O = (0, 1)$ (0% false positive, 100% true positive). The line $y = x$ includes the default hypotheses: at one extreme (TP=0, FP=0), the default hypothesis puts all examples in the negative class, at the other extreme (TP=1,FP=1), it puts all examples in the positive class.

Let r denote the ratio of false negative vs false positive error cost; then, the best possible result is found at the intersection of the ROC curve with the Δ line, starting from the optimal point O and having slope $-\frac{1}{r}$. One sees from Fig 1 that which one is the best learning algorithm, depends on the r ratio.

4.2 Stability requirements

Two requirements are made for supervised learning from medical data.

Firstly, the desired algorithm should allow for inexpensive tuning of the error costs; this way, the expert can efficiently choose the best trade-off between the false positive and false negative error rate. Practically, the learning algorithm should provide the medical expert with the so-called ROC curve, and allow him/her to navigate this curve and decide which range of sensibility (true positive rate) and specificity (true negative rate) is worth.

Secondly, this algorithm should achieve stable learning, ie the hypotheses should "vary gracefully" as the error costs are modified. This not only means that hypotheses with similar behavior should be retrieved as the error costs vary; ideally, the hypotheses should be syntactically as similar as possible.

As will be seen in the next section, these requirements are addressed as the main induction step is made independent from the error costs. The learned hypothesis is minimally adjusted thereafter depending on the error costs.

5 Learning with AUC

This section describes and discusses a novel learning algorithm, termed *EROL* for Evolutionary ROC-based Learning.

5.1 Principle

Practically, the hypothesis space $\mathcal{H}$ considered in the following is included in the space of real-valued classifiers, mapping the instance space X onto the real-valued space $\mathbb{R}$. Such a classifier h induces by thresholding a family of binary classifiers $\{h_t,\ t \in \mathbb{R}\}$, with $h_t(x) = 1$ if $h(x) > t$ and -1 otherwise.

It is straightforward to see that the true positive and the false positive rates monotonically increase as t decreases. Therefore there exists a single optimal value for t, optimizing the generalization error for a given cost matrix.

Within this framework, the induction algorithm focuses on learning the real-valued hypothesis h; the cost matrix is accounted for by only adjusting the threshold t. This adjustment effectively ensures the stability of the hypotheses produced wrt the misclassification costs.

The overall assessment of the real-valued hypothesis h is based on the ROC curve associated to the set of binary classifiers h_t, for t ranging in $\mathbb{R}$.

It comes naturally to assess hypothesis h from the area under this ROC curve (AUC). Actually, h is constructed by optimizing the AUC criterion.

Indeed, this approach is strongly inspired from the margin-based learning approaches such as SVMs, bagging [2] or boosting [8], which similarly lie in the space of real-valued classifiers. The difference between *EROL* and SVMs lie in the underlying criterion; the difference between *EROL* and comittee learning, is that the core learning process in *EROL* relies on optimization.

5.2 *EROL*

The AUC criterion induces a mixed optimization problem, for the area under the ROC curve depends on the order permutation σ on the dataset, which is induced by the numerical function h.

This optimization problem is therefore tackled using evolutionary computation techniques, which are population-based stochastic optimization algorithms, crudely inspired from the natural evolution of biological populations and Darwinian principles [4, 1].

Assuming the reader's familiarity with canonical GAs [4], we only detail *EROL* specificities for the sake of reproducibility.

The fitness function computing the AUC criterion is described in Table 4 (normalization is omitted as it has no effect on the optimization problem).

Fitness function of hypothesis h

```
Input
  Data set $\mathcal{E} = \{(x_i, y_i), i = 1 \ldots n,\ x_i \in X, y_i \in \{1, -1\}\}$
  Hypothesis $h : X \mapsto \mathbb{R}$
Init
  Sort $\mathcal{E} = \{(x_i, y_i)\}$ by decreasing order, where $i > j$
    iff $(h(x_i) > h(x_j))$ or $((h(x_i) = h(x_j)$ and $(y_i > y_j))$.
  $p = 0$
  $\mathcal{F} = 0$
For $i = 1$ to $n$
  if $y_i = 1$, increment $p$;
  else $\mathcal{F} = \mathcal{F} + p$
EndFor
Return $\mathcal{F}$
```

Table 4. Fitness of h = Area Under the Roc Curve of h

We restrict ourselves to learning in attribute-value logic; accordingly, the instance space X is set to $\mathbb{R}^p$, where the nominal attributes with k modalities are handled as k boolean attributes.

The hypothesis search space $\mathcal{H}$ considered in this paper is the set of linear functions on X. A linear hypothesis h is defined as a real-valued vector belonging to $\mathbb{R}^p$, still noted h by abuse of notation.

The optimisation of the fitness function $\mathcal{F}$ on the search space $\mathcal{H} = \mathbb{R}^p$ is performed using evolution strategies (ES).The main difference compared to GAs regards the mutation operator, specifically tailored to numerical optimization. Mutation in ES is used much more intensively used than in GAs, and will be used with probability 1 in the experiments. The mutation of the i-th component of individual h is done by addition of a Gaussian perturbation $\mathcal{N}(0, \sigma_i)$. We used self-adaptive mutation, where the genetic material of the individual incorporates the standard deviations σs; this way, evolution adjusts the mutation amplitude depending on the generation and the individual. A genetic individual thus is a vector in $\mathbb{R}^{2p}$. The reader is referred to [1] for an exhaustive presentation of self adaptive mutation.

We use the $(\mu+\lambda)-ES$ selection/replacement mechanism; μ parents generate λ offspring, and the best individuals among the μ parents + λ offspring are selected as parents for the next generation.

Note that *EROL* is not restricted to linear hypotheses. The kernel trick popularized from SVMs allows *EROL* for building more complex hypotheses with little additional cost. This is achieved by exploring the mixed search space $\mathcal{H} = (\mathbb{N} \times \mathbb{R})^v$, where a hypothesis h is characterized from a subset of training examples, with indices $i_1 \ldots i_v$ and their associated weights $\alpha_1 \ldots \alpha_v$. Hypothesis h is here defined as

$$h(x) = \sum_{k=1}^{v} \alpha_k K(x_{i_k}, x)$$

6 Validation

We used with *EROL* the same data as in the Bivaried Analysis. Two groups of patients belonging to the risk group. One group contains all the healthy patients at the end of the medical campaign. The other one contains the ill or dead patients.

We have studied three different groups of attributes and built three different tests with them. Firstly, the whole set of attributes previously study (see section 3.1 for details).

Secondly, the 'free of charge' attributes (with no need of medical intervention and without blood test) namely: age, bmi, tobacco, alcohol, physical activities, marital status and educational level.

Thirdly, the last test was built with all the medical attributes (blood test and blood pressure), including bmi and age.

These tests were learned using *EROL*$\dot{\text{D}}$atasets has been divided into 1/3 for learning and 2/3 for test. We ran 20 differents runs and compute the median result. We have tried different parameters for the fitness fonction of *EROL*

- learning an additive or **multiplicative** fonction
- using negative weigths versus **not negative**

The bold parameters gives the best results for our problem.

The Figure 2 present the three medical tests.

For the best test (the third one) we compute the Predictive Positive Value[2]. The Predictive Positive Value is the probability to be ill if the test is positif. This value is equal to 90.1%.

So using this test on a risk population, the pratician is more sure that ill patients are really ill, and it is easier for him to well treat them.

[2] http://cri-cirs-wnts.univ-lyon1.fr/polycopies/cardiologie/cardiologie-110.html

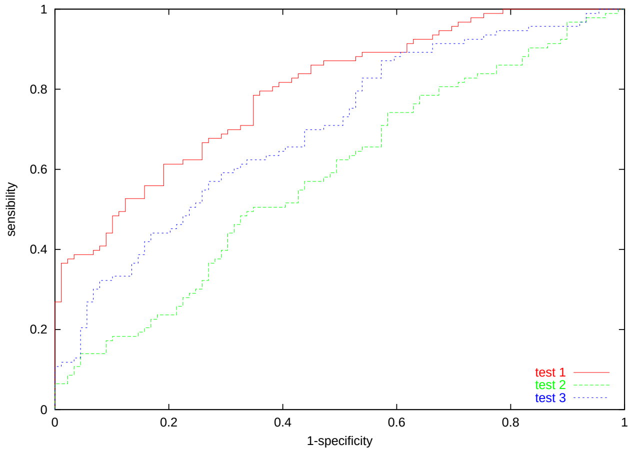


Fig. 2. The medical tests.

We have applied this test to the whole *Entry* dataset in order to compute its sensibility[3] and specificity[4] values.

To compute this two values, we used the following protocole. We considered as reference the individual in the low risk class, having maximal risk estimate. This individual truly belongs to the low risk class, for he also appears in the *Control* database and did not present any atherosclerosis-related symptom during the whole study duration. We then considered the subset of individuals classified in the medium risk class, normal class, or unknown class, whose rik estimate is greater than that of the avove reference individual.

This subset includes 199 individuals, among whom 135 appear in the *Control* database. The patients were distribute as indicate in the Figure 3.

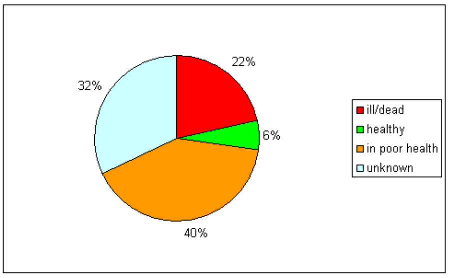


Fig. 3. Patients distribution.

[3] capacity for a test to identify ill patients

[4] capacity for a test to identify healthy patients

This test has a sensibility of 82.69% and a specificity of 93.45%. It can be used to help the physician to decide more examinations for a patient. More, this test is very cheap (only a blood test and a medical visit).

7 Conclusion

This investigation into the aggravating factors for cardiovascular diseases has been deeply informative, at least for us, for several reasons.

First of all, it has confirmed some general statements about the causes of CVD e.g. typically, tobacco intoxication is harmful. It has also shown the limitations of automatic exploratory analysis; this was exemplified on the first results obtained regarding the impact of sport practice. As the result was not acceptable due to our background knowledge, we were led to reconsider our descriptions of the problem; this redescription duly resulted in showing the beneficial influence of leisure activities, which was previously unnoticed and hidden by other physical activities. Further research perspectives include firstly the refinement of the risk estimate, based on optimization techniques in computer sciences and secondly the test enrichment with the most recent discoveries in medical research. For example, a recent work placed C-Reactive Protein [3], not knew in the seventies, in the leaders factors for cause of atherosclerosis and thrombosis.

References

1. T. Bäck. *Evolutionary Algorithms in theory and practice.* New-York:Oxford University Press, 1995.
2. L. Breiman. Arcing classifiers. *Annals of Statistics*, 26(3):801–845, 1998.
3. P. Libby et al. Inflammation and atherosclerosis. *Circulation*, 105(9):1135–1143, 2002.
4. D. E. Goldberg. *Genetic algorithms in search, optimization and machine learning.* Addison Wesley, 1989.
5. D.J. Hand and R.J. Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine Learning*, 45(2):171–186, 2001.
6. P. Langley. Data-driven discovery of physical laws. *Cognitive Science*, 5:31–54, 1981.
7. N. Lucas, J. Azé, and M. Sebag. Atherosclerosis risk identification and visual analysis. In *Discovery Challenge ECML-PKDD 2002.* http://lisp.vse.cz/challenge/ecmlpkdd2002/, 2002.
8. R.E. Shapire, Y. Freund, P.Bartlett, and W.S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. In *Proceedings of the 14th International Conference on Machine Learning*, pages 322–330. Morgan Kaufmann, 1997.
9. S. Siegel. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.
10. D. Wettschereck. Educational data preprocessing. In *Discovery Challenge ECML-PKDD 2002.* http://lisp.vse.cz/challenge/ecmlpkdd2002/, 2002.