

Atherosclerosis Risk Identification and Visual Analysis (PKDD 2002 Challenge)

Noël Lucas¹, Jérôme Azé², Michèle Sebag²

(1) Lab. CEDRIC, CNAM, F-75141 Paris

(2) LRI, CNRS UMR 8623, Université d'Orsay, F-91405 Orsay

Abstract. This paper is concerned with the analysis of atherosclerosis factors. The main originality of the approach is to handle risk identification as a regression problem instead of a classification problem. Arguably, what makes sense is considering that a given individual is more at risk than another one; deciding among high and medium risk involves many other considerations (e.g. socio-economical) than medical ones.

Practically, a real-valued risk estimate is proposed. This estimate allows for questioning the classification of individuals in the available data, and exploiting the control data for coherence checking.

Additionally, a visual sensitivity analysis of cardiovascular risk wrt the main aggravating factors, e.g. tobacco and alcohol intoxication, sport activities and overweight, based on this estimate, is presented and discussed.

1 Introduction

A wealth of information has been gathered for several decades in hospitals and during medical campaigns. The available data describes patients (sex, age, weight,...), the interventions/prescriptions/diseases they have undergone, and the examination results. The exploration and exploitation of these data aim at three goals:

- detect the coming diseases (or acute crises) for their better prevention;
- ensure that invasive or expensive interventions/exams are prescribed when needed and only then;
- bring some understanding of the causes and side-effects entangled in the disease phenomenon.

The paper is concerned with a key challenge for Public Health, the detection and prevention of cardiovascular diseases. These diseases have in common the fact that the cells in the walls of the arteries are severely damaged. All such damages constitute atherosclerosis; they develop for a long time before pain and other symptoms manifest. The prevention of atherosclerosis thus is undoubtedly among the most touching concerns for medical data mining, besides its heavy socio-economical implications.

The available data have been collected at IKEM (Praga) and the Medicine Faculty at Charles University (Plzen), and were kindly provided as Challenge for PKDD02.

These data concern the risk of atherosclerosis and cardiovascular health; depending on personal and family history, the individuals have been divided into three classes, corresponding to low, medium and high cardiovascular risk.

The goal of the study is to inspect the relationships between the various known aggravating factors for cardiovascular diseases (tobacco, alcohol, overweight, physical activities) and see whether and how these relationships are modified in the various risk groups¹. Assuming that the study ultimate goal is to provide support for Health-Policy Making, the paper aims at enabling a visual inspection of these relationships. The approach, inspired from [3], puts the stress on presenting the information in graphical and simple terms to the expert, allowing him/her to use *vision plus background knowledge to think and decide*.

Section 2 briefly describes the data, and presents our goal. Section 3 details the data preparation which was performed. Sections 4 and 5 report the results obtained and discuss their relevance from a medical viewpoint. We last conclude on the perspectives for medical data mining, opened by the approach.

2 Data and goals

The present work mostly focuses on the ENTRY database provided in the PKDD02 Challenge. This database reports on 1,419 patients divided in three classes, corresponding to low, medium and high risk profiles.

Various types of information have been collected on the patients. This information illustrates six groups of risk factors and indicators:

- Observable factors include the laboratory examinations (glycemy, cholesterol, moc, triglycerid, hypertension, fold skins, weight, height) enriched with compound attributes (inspired from body-mass index, BMI).
- Personal case history factors include previous attacks of myocardis, arterial hypertension, diabetes, ictus, hyperlipidemy.
- Family factors include the family case history, e.g. first degree parents' death due to cardiovascular-related diseases, presence of myocardis, diabetes, etc. and sudden death in the patient family.
- Socio-economical factors include the education level, marital status, professional responsibility.
- Habitus factors describe the individual way of life (does he allegedly smoke, drink, practice sport and so forth). Clearly, these factors can hardly be considered as observable factors, for their appreciation might be subjective in some cases, and their reliability might be subject to caution in other cases.
- Factors related to pain and comfort include angor and intermittent claudication.

2.1 Objectives, goals of experiments

Our goal is twofold. First, we want to predict the patient risk depending on all information available. In particular, the risk function to be learned must accommodate the fact that only part of the medical factors might be known at examination time.

¹ Technically, this goal would be amenable to learning and comparing joint distributions [6].

To this aim, two partial risk functions are constructed independently, respectively measuring the risk related to the family case history, and the risk related to the personal case history.

These two risk functions are combined to form the global risk estimate, which is validated against the available database. Interestingly, the validation process opens on investigating anew the data, checking whether the patients who have been classified in a given risk group might have been misclassified.

The available data offer unusual opportunities to question the initial classification of the individuals, for an additional database (CONTROL) reports on the medical history of a sample of the patients over some years. In retrospect, some differences between patients who were initially classified in the same class might thus be justified and argued.

A second goal is to provide the physician, or rather the health policy maker, with a visual tool for investigating the effects of the various factors (e.g. tobacco, alcohol, overweight, physical activities) on the risk. As emphasized in [3], human eyes are among the most sophisticated tools for synthesizing the global tendencies in complex phenomena. The only requisite for *using vision to think* is to provide the expert with a 2D or 3D representation of the phenomenon under study, such that this representation is endowed with a self explanatory semantic, and the representation biases are amply documented.

Let us first describe the data preparation phase.

3 Data Preparation

As widely acknowledged in the Data Mining literature [5], data preparation is the key step for the success of mining and learning algorithms.

Data have thus been prepared to form a reduced number of variables.

3.1 Family case history

The family case history raises the following difficulty. The number of available variables (219) is very large compared to the number of patients; accordingly, these variables are sparingly informed in the database. We therefore create aggregate variables, as an attempt to compress the available information in the database. Needless to say the choice done is subjective on one hand, and depends on the data on the other hand.

- a single boolean variable accounts for all variables related to the father's death (before 55) and the mother's death (before 65) due to IM, ictus or sudden death (disjunction).
- likewise, a boolean variable accounts for all variables related to the father's death (before 55) and the mother's death (before 65) due to Diabetes, HTA or Angor (disjunction).
- a boolean variable accounts for all variables related to the father's death (between 55 and 65) due to IM, ictus, HTA, Diabetes, Angor, NAHL.

- a boolean variable accounts for having (at least) one brother or sister affected by Ictus, IM or sudden death (disjunction).
- one boolean variable accounts for the aggravated diabetes in the family (both parents are alive and suffer from diabetes)
- one boolean variable accounts for moderately aggravated diabetes (the father or mother, and one brother or sister, are alive and suffer from diabetes).
- one boolean variable accounts for moderately aggravated angor (the father or mother, and one brother or sister, are alive and suffer from angor).
- one boolean variable accounts for moderately aggravated HTA (the father or mother, and one brother or sister, are alive and suffer from HTA).
- one boolean variable accounts for moderate family risk (at least two first degree parents (father, mother, brothers, sisters) are alive and suffer from IM or HTA or Angor or Ictus, or Diabetes).

3.2 Personal case history, A2 questionnaire and biochemical exams

The goal is to estimate the atherosclerosis risk from the patient's age, his symptoms or diseases and the way these diseases have been combated through diet or medication.

- One variable simply is the patient's age minus 15 (assuming that the arteries state was perfect at that time).
- One variable is the number of years where the patient has been suffering from HTA or hyperlipidemy; this number is multiplied by two if the patient did not follow the physician's prescription;
- One variable is the number of years where the patient has been suffering from Diabetes; similarly, this number is multiplied by two if the patient was not following his diet.
- One variable reflects the number of acute symptoms such as IM, Ictus, Angor, intermittent claudication, acute asthma (grade IV) or albumin.
- One variable reflects the presence of asthma (grade II or III).
- One variable reflects the presence of sugar in moc for patients who are not known as diabetic, as this symptom is a precursor for diabetes.
- One variable reflects the presence of cholesterol or triglycerides (greater or equal the reference threshold 250).

These variables typically are taken into account through an exponential law as the damage to the cells in the walls of the arteries increases fast as the arteries are damaged [4].

3.3 Toxicological problems

Two real-valued variables have been considered, measuring 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 cardiovascular 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 smoked (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 modeled through a multiplicative factor. The tobacco intoxication factor is multiplied by .8 if the interruption started less than one year ago and by .8 otherwise.

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

Another variable measures the energy spent daily in leisure activities. The distinction between global energy and leisure activities will be argued in section 5.4.

4 The risk estimate: an expert approach

This section describes the manual exploitation of the variables defined in the previous section.

4.1 Rationale

Our claim is that the scientific difficulty does not regard the characterization of extreme cases. Individuals with healthy life and no family case will almost surely have a low risk; symmetrically, heavy drinkers and/or smokers, overweighted and practicing no sport, will have a high risk.

The actual challenge is to decide about other individuals, who are neither very healthy, nor heavily loaded. Typically, the boundary between low and medium cardiovascular risk governs the prevention policy; nothing should be done for low risk individuals, while some efforts might be devoted to medium risk individuals.

Symmetrically, the boundary between medium and high cardiovascular risk also governs the prevention policy (prevention is necessary for high risk individuals, and probably useful but not to the same degree for medium risk individuals).

Along these lines, we suggest that cardiovascular risk is a relative (ordered) concept more than an absolute (nominal) one. What makes sense is the fact that a given individual is more at risk than another one, as opposed to being at risk *per se*.

Therefore, we formalize the notion of risk as a real-valued function instead of a nominal one. This way, the point of setting thresholds (e.g. determining the boundary between high and medium risk) can soundly be delegated to optimization-oriented techniques, relying on more informed analysis (taking into account the prevention undesirable side effects, the economic issues, and so forth).

4.2 The risk estimate

Accordingly, the underlying classification problem (the patients being divided into three classes) is rather addressed as a regression problem. We defined a global risk estimate as the sum of

- A linear combination of the variables related to the personal case history; this part corresponds to the static part of the cardiovascular risk (heredity-related health capital).
- An exponential function of the variables related to the personal case history; as mentioned earlier, the cardiovascular diseases effects are cumulative.

The coefficients involved in the global risk estimate have been adjusted manually after a few preliminary experiments. All in all, five distinct coefficients have been adjusted².

On-going experiments are concerned with optimizing the coefficients using GAs [2]. As is, the risk estimate can be graphically compared wrt the initial risk. Fig. 1 displays the individuals sorted by increasing estimated risk; for each individual is reminded his initial risk class (values 5, 3 and 1 respectively correspond to high, medium and low risk).

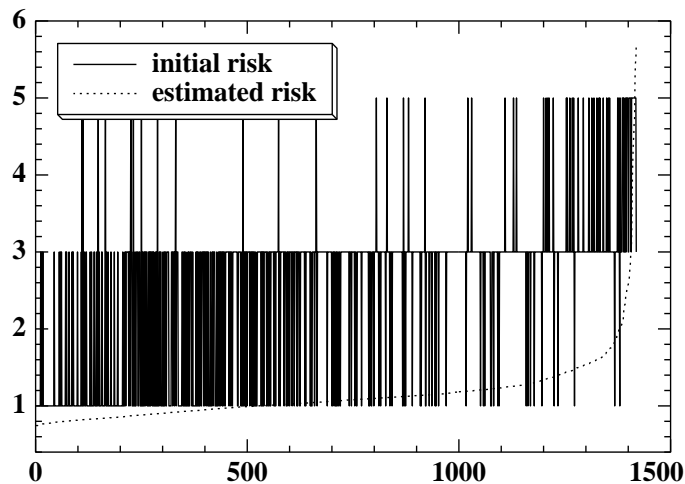


Fig. 1. Initial risk vs Risk estimate

As the high risk class actually includes patients who already suffer from atherosclerosis, the key validation criterion is to distinguish among the low and medium risk classes; the challenge is to detect the patients who would best benefit from preventive health care.

² Due to space limitations, all additional details related to the data preparation, including the data themselves, will be found at <http://www.lri.fr/~aze/PDDD2002>.

4.3 Validation and case analysis

The comparison between the risk estimate and the initial risk shows significant differences. A case analysis was thus performed in order to understand these differences.

We considered as reference the individual in the low risk class, having maximal risk estimate (1.8). This individual (reference 20300) 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, whose risk estimate is greater than that of the above reference individual. This subset includes 24 individuals, among whom 22 appear in the CONTROL database³. Suggestively, all 22 individuals became seriously ill and all (except 2 of them) suddenly disappeared from the CONTROL database (one plausible explanation was that they were dead, though nothing in the database could infirm or confirm this conjecture).

These 22 individuals were compared to the average individual in the medium risk class, and they show significantly different (according to a Chi2 test with significance level 1%) in three respects.

First of all, these individuals are more overweighted than other individuals in class 3; however, this was not unexpected, since these individuals score high on our risk estimate and the overweight factor duly intervenes in this estimate.

What is more convincing, they significantly differ from other individuals in class 3 wrt atherosclerosis-originated diseases. A boolean variable, coding for one or more among (IM, Angor, MI, cerebro-vascular accident, ischemic heart disease, claudication, silent myocardial ischemia, silent myocardial infarction) was computed for all individuals with medium risk in the CONTROL database; again, the 22 individuals differ from the others according to a Chi2 test with significance level 1%.

Last, the difference is even more striking (significance level < 1%) when diabetes is taken into account besides the atherosclerosis effects.

The partial relevance of the risk estimate proposed above can thus be argued, as being (at least) complementary wrt the risk score used in the study. After this estimate, these 24 individuals would have been detected as being at risk, and they would therefore have benefit from specific preventive health care, avoiding or delaying cardiovascular accidents and pain. Among the perspectives for further research is the design of specific tests, characterizing such high risk profiles.

Symmetrically, on-going case study is concerned with individuals classified in the medium risk class, who inversely score low after our estimate.

4.4 Discussion

We are aware of the fact that a great many real-valued estimates would have been compatible with the risk groups in the data. The main justification lies in the simplicity of the model on one hand, and its conformity with the current body of knowledge in medicine (see e.g. [4]).

³ The individual references are 10030; 10044; 10115; 10309; 10384; 10396; 10404; 10441; 10502; 10557; 10594; 10728; 10750; 10847; 10851; 10865; 10914; 10920; 10958; 10963; 10988; 10996.

Among the limitations of this estimate is the fact that several factors which have been shown critical for atherosclerosis in the last decade, were ignored at the study time and are thus missing in the database. Such is the case of inflammatory factors (CRP test, C-reactive protein) that govern the triggering and development of atherosclerosis.

5 Sensitivity analysis

As mentioned in the introduction, we understood the central challenge goal as studying the relationships between the various factors for cardiovascular diseases, and determining whether these relationships vary among the various risk groups. A natural approach would have been to estimate separate joint distributions on the various groups, and compare these distributions (see, e.g. [6]); the confidence on such comparisons clearly measures the representativity and quality of the data.

Rather than automatically learning and comparing joint distributions, we provide the expert user with visual elements for inspecting and comparing the impact of the cardiovascular factors. After detailing the principle, the results obtained regarding the four aggravating factors (alcohol, tobacco, overweight and (lack of) physical activity) are displayed and interpreted.

5.1 Principle

The approach is based on the risk estimate presented in section 4. The general idea is that enriching the risk (as a real-valued function instead of a nominal one) will allow the expert to make fine-grained differences.

Practically, each considered aggravating factor is measured as a real-valued attribute. We thus select the first group of individuals as being the 5% individuals with minimal values for the factor at hand; the second group symmetrically includes the 5% individuals with maximal values for the factor.

As a first attempt, the risk estimate of these two individual groups was displayed (Fig 2), where the individuals are ranked according to the factor under study. However, this was not satisfactory as the resulting curve was difficult to read.

A second attempt was done, where individuals are ranked with increasing risk estimate (Fig 3, a).

Though much more legible, this representation does not convey the relation between the risk and the aggravating factor at hand; typically, one does not know whether those individuals with high risk are uniformly distributed among the smokers, or concentrated among the very heavy smokers.

A third attempt (Fig. 3, b) thus displays the cumulative risk: individuals are ranked according to the factor at hand; at coordinate $x = i$ is associated the cumulated risk $y = \sum_{j=0}^i risk[j]$. A linear curve with same slope and intercept ($y = x[0] \times x$) shows that the risk hardly depends on the factor at hand. Inversely an exponential-like curve shows that the risk steadily increases depending on the factor at hand.

We finally decide to propose both the cumulative risk and risk-based ranking curves to the expert to support his/her interpretation.

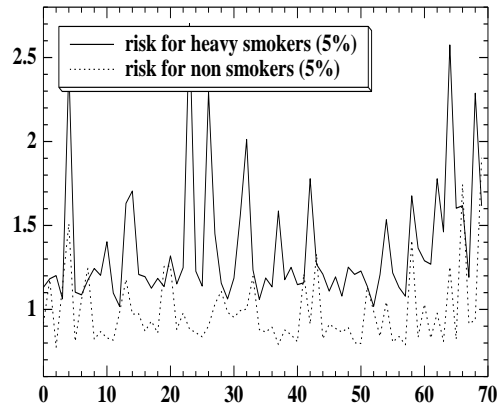


Fig. 2. Influence of Tobacco on cardiovascular risk

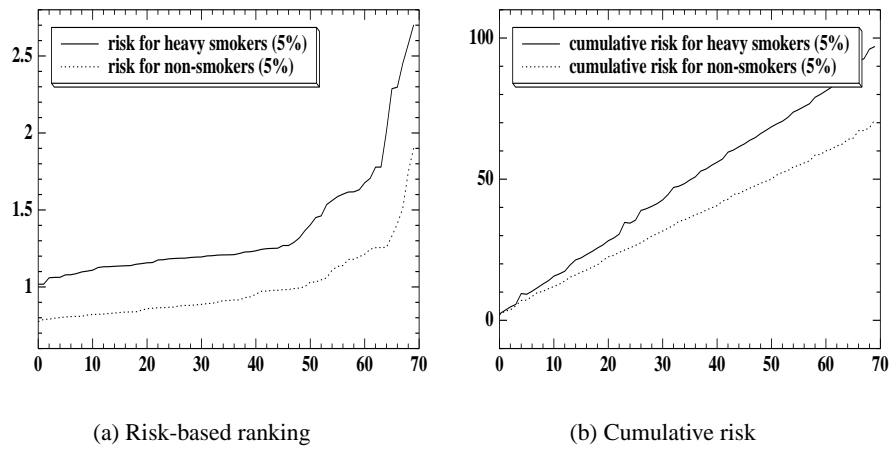


Fig. 3. Influence of Tobacco on cardiovascular risk

5.2 Tobacco impact

The above discussion was illustrated on the curves representing the tobacco impact on cardiovascular risk.

As could have been expected, tobacco is an aggravating factor for cardiovascular diseases; this is manifest as the risk estimated for the heavy smokers is significantly and consistently higher than for non smokers. According to our estimate, the one third with lowest risk scores below 1 (including more than 2/3 of the non-smokers); the one third with highest risk scores above 1.1 (including more than 9/10 of the heavy smokers).

5.3 Alcohol impact

Unexpectedly, the impact of alcohol appears significantly less harmful than that of tobacco on cardiovascular risk (Fig 4). In retrospect, the major difference is that non-drinkers do not score very low (compared to non-smokers); this might be related to the fact that drinking with moderation (especially wine) allegedly has a beneficial impact wrt cardiovascular diseases.

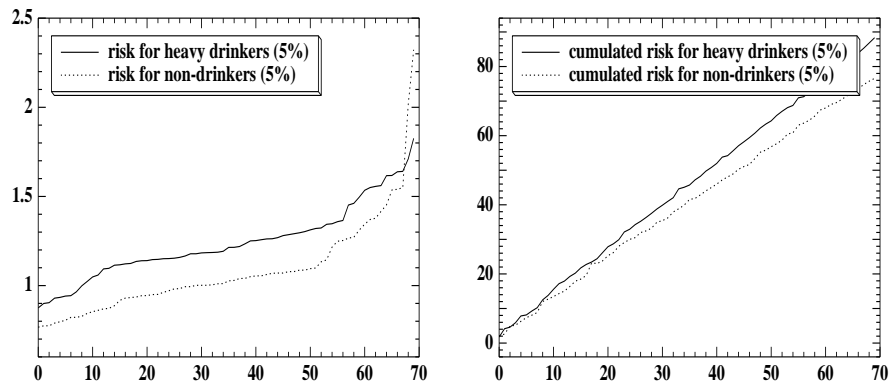


Fig. 4. Influence of Alcohol on cardiovascular risk

5.4 Physical activities and leisure

In a first step, we consider as relevant factor the total energy daily spent by the individual (at work, during transports, and at sport). What was indeed unexpected, the impact of physical activity then appeared insignificant (during the first half of the curve) and even harmful (during the second half of the curve) (Fig. 5). As we were not prepared to accept this conclusion, we reconsidered the definition of the variable (section 3.4), and made the distinction between energy spent for work and in leisure activities. The beneficial influence of sportive activities then became apparent (Fig 6). In retrospect, this beneficial influence was hidden as very few kcal are spent in leisure activities compared to work and transportation.

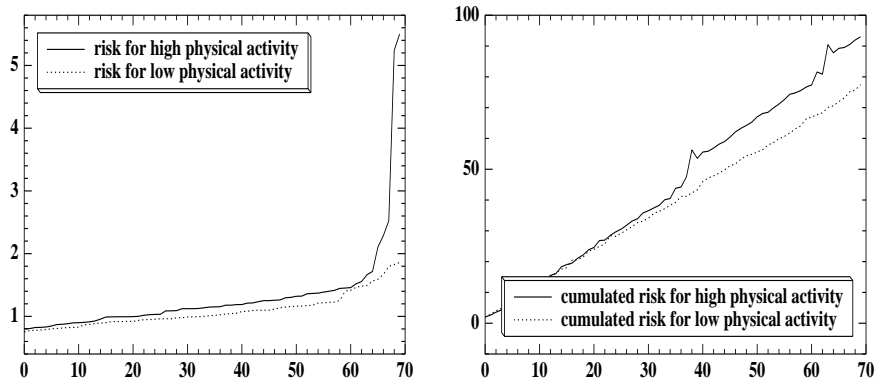


Fig. 5. Influence of Physical Activities on cardiovascular risk

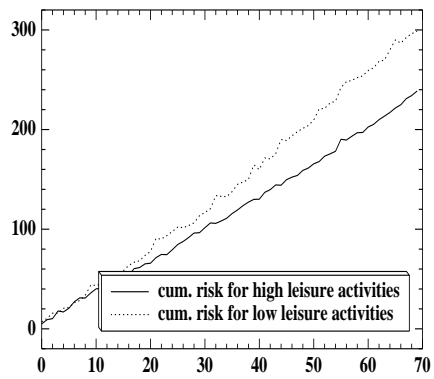


Fig. 6. Influence of Leisure Activities on cardiovascular risk

5.5 Overweight factor

Similar curves related to the impact of body mass index (not displayed for space limitations) suggest that extreme values for BMI are harmful in what regards cardiovascular diseases. Case analyzes are underway to investigate this phenomenon in more depth.

6 Conclusion and perspectives

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 description 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 the refinement of the risk estimate, based on optimization techniques. Genetic algorithms [2] and Genetic programming [1] will be applied to parametric and non parametric identification of a satisfactory risk estimate. The key advantage of evolutionary computation is to accommodate the fact that the learning goal is actually under-constrained (using classified data for tackling a regression problem) on one hand, and incorporate additional, knowledge-based constraints.

References

1. W. Banzhaf, P. Nordin, R.E. Keller, and F.D. Francone. *Genetic Programming — An Introduction On the Automatic Evolution of Computer Programs and Its Applications*. Morgan Kaufmann, 1998.
2. T. Bäck. *Evolutionary Algorithms in theory and practice*. New-York:Oxford University Press, 1995.
3. S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Information Visualization: Using vision to think*. Morgan Kaufmann, 1999.
4. P. Libby et al. Inflammation and atherosclerosis. *Circulation*, 105(9):1135–1143, 2002.
5. U.M. Fayyad and B.K. Irani. Multi-interval discretization of continuous valued attributes for classification learning. In *Proc. of IJCAI'93*, pages 1022–1027. Morgan Kaufmann, 1993.
6. J.M. Pena, J.A. Lozano, and P. Larranaga. Learning recursive bayesian multinets for data clustering by means of constructive induction. *Machine Learning*, 47:63–90, 2002.