Analysing and Improving the Diagnosis of Ischaemic Heart Disease with Machine Learning

Matjaž Kukar[1], Igor Kononenko[1],
Ciril Grošelj[2], Katarina Kralj[1], Jure Fettich[2],
[1]University of Ljubljana
Faculty of Computer and Information Science,
Tržaška 25, SI-1001 Ljubljana, Slovenia
tel/fax: +386 61 1768 386
[2] University Medical Centre Ljubljana,
Nuclear Medicine Department,
Zaloška 7, SI-1001 Ljubljana, Slovenia
e-mail: {matjaz.kukar, igor.kononenko}@fri.uni-lj.si

Key words: Machine Learning, Ishaemic heart disease, Cost-sensitive learning, ROC analysis, Feature subset selection

Abstract

Ischaemic heart disease is one of the world's most important causes of mortality, so improvements and rationalization of diagnostic procedures would be very useful. The four diagnostic levels consist of evaluation of signs and symptoms of the disease and ECG (electrocardiogram) at rest, sequential ECG testing during the controlled exercise, myocardial scintigraphy, and finally coronary angiography (which is considered to be the reference method).

Machine Learning methods may enable objective interpretation of all available results for the same patient and in this way may increase the diagnostic accuracy of each step. We conducted many experiments with various learning algorithms and achieved the performance level comparable to that of clinicians. We also extended the algorithms to deal with non-uniform misclassification costs in order to perform ROC analysis and control the trade-off between sensitivity and specificity. The ROC analysis shows significant improvements of sensitivity and specificity compared to the performance of the clinicians. We further compare the predictive power of standard tests with that of Machine Learning techniques and show that it can be significantly improved in this way.

1 Introduction

Ischaemic heart disease (IHD) is the most important cause of mortality in developed as well as in developing countries. Therefore improvements and rationalization of diagnostic procedures and treatment of IHD are necessary.

The usual procedure in IHD diagnosis consists of four diagnostic levels, which contain evaluation of signs and symptoms of the disease and ECG at rest, sequential ECG testing during a controlled exercise, myocardial scintigraphy and coronary angiography as a final test. Because suggestibility is possible, the results of each step are interpreted individually and only the results of the highest step are valid. The total amount of data available for each patient is too large to be efficiently and objectively evaluated by the clinicians.

The goal of a rational diagnostic algorithm is to establish the conclusive diagnosis of IHD and to plan the most appropriate management of the disease using only the necessary diagnostic steps. This can be achieved by taking into account and evaluating all the information collected by different diagnostic methods according to their importance and diagnostic value.

The performance of a diagnostic method is usually described as classification accuracy, sensitivity and specificity:

$$accuracy = \frac{\#true\ positives + \#true\ negatives}{\#all\ patients} \quad (1)$$

$$sensitivity = \frac{\#true\ positives}{\#all\ patients\ with\ the\ disease} \quad (2)$$

$$specificity = \frac{\#true\ negatives}{\#all\ patients\ without\ the\ disease} \quad (3)$$

The *true positives* are all patients with the disease and positive test result, whereas the *true negatives* are all patients without the disease and negative test result.

The reported average values of these measures, taken from 29 reports containing several thousands of patients are as follows [Gerson, 1987]. Sensitivity for the exercise ECG (5796 patients) is 72%, specificity 79%, and accuracy 74%. For the myocardial scintigraphy (2413 patients) they are 84%, 88%, and 85%, respectively. In both cases the coronary angiography is a reference method.

The aim of this study is to improve the diagnostic performance (sensitivity and specificity) of non-invasive diagnostic methods (i.e. clinical examinations of the patients, exercise ECG testing, and myocardial scintigraphy in comparison with the coronary angiography as a definite proof of coronary artery stenosis) by evaluating all available diagnostic information with Machine Learning techniques. The ultimate goal was to reduce the number of patients that must unnecessarily be submitted to further invasive pre-operative examinations (these can be potentially dangerous, unpleasant and very costly).

The paper is organized as follows. Section 2 describes the usual diagnostic process and

the data that were used in our experiments. Section 3 briefly describes the applied Machine Learning algorithms (classifiers). We have extended several Machine Learning algorithms to take into account the misclassification costs. Section 4 describes the extensions. In Section 5 we compare the performance of the classifiers in terms of prediction accuracy, information score, sensitivity and specificity against the clinical results, as well as the usefulness of cost-sensitive learning for the ROC analysis. In Section 6 we show how the predictive power of the tests can be further improved by using the Machine Learning techniques. In Section 7 we discuss the results and suggest possibilities for further work.

2 The Diagnostic Problem and the Dataset

The function of the heart is to pump blood to all organs of the body. For this task an uninterrupted and continuous supply of oxygen to the heart muscle is needed. This is achieved by sufficient blood flow through the coronary arteries to the heart muscle – myocardium. In case of diminished blood flow through coronary arteries due to stenosis or occlusion, IHD develops, producing impaired function of the heart and finally the necrosis of the myocardium – myocardial infarction.

During the exercise the volume of the blood pumped to the body per minute has to be increased and therefore the delivery of the oxygen to the heart muscle has to increase several times by increasing the blood flow trough the coronary arteries. In a (low grade) IHD the blood flow as perfusion of the myocardium is adequate at rest or during a moderate exercise, but insufficient during a severe exercise. Therefore, signs and symptoms of the disease develop only during the exercise.

There are four levels of diagnostics of IHD. Firstly, signs and symptoms of the disease are evaluated clinically and ECG is performed at rest. This is followed by sequential ECG testing during controlled exercises by gradually increasing the work load of the patient. Usually a bicycle ergometer or treadmill is used to establish the diagnosis of IHD by evaluating changes of ECG during the exercise (Figure 1).

If this test is not conclusive, or if additional information regarding the perfusion of the myocardium is needed, myocardial scintigraphy is performed. Radioactive material is injected into the patient during exercise. Its accumulation in the heart is proportional to the heart's perfusion and can be shown in appropriate images (scintigrams). Scintigraphy is repeated at rest and by comparing both sets of images, the presence, the localization, and the distribution of the ischaemic tissue are determined (Figure 2).

If an invasive therapy of the disease is contemplated, i.e. the dilatation of the stenosed coronary artery or coronary artery bypass surgery, the diagnosis has to be confirmed by imaging of the coronary vessels. This is performed by injecting radio opaque (contrast) material into the

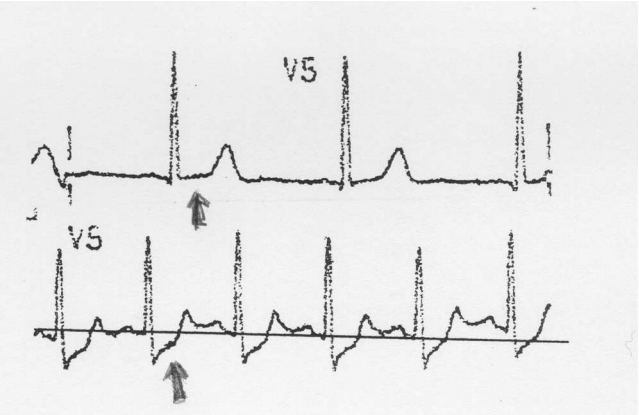

Figure 1: Positive test result of the exercise ECG. The downslopping of the so-called ST line shows the presence of IHD.

coronary vessels and by imaging their anatomy with x-ray coronary angiography (Figures 3 and 4).

In our study we used a dataset of 327 patients (250 males, 77 females) with performed clinical and laboratory examinations, exercise ECG, myocardial scintigraphy and coronary angiography because of suspected IHD. The features from the ECG an scintigraphy data were extracted manually by the clinicians. In 229 cases the disease was angiographically confirmed and in 98 cases it was excluded. 162 patients had suffered from recent myocardial infarction. The patients were selected from a population of approximately 4000 patients who were examined at the Nuclear Medicine Department between 1991 and 1994. We selected only the patients with complete diagnostic procedures (all four levels) [Kukar et al., 1997]. Our experiments were conducted on four problems. They differ in the amount of clinical and laboratory data (attributes) available for learning, corresponding to different diagnostic levels (Table 1).

Clinical and laboratory tests performed	Number of attributes		
	Nominal	Numeric	Total
Signs and symptoms	23	7	30
Signs, symptoms and exercise ECG	30	16	46
Signs, symptoms, exercise ECG and scintigraphy	52	25	77
Myocardial scintigraphy only	22	9	31
Entropy of the dataset	0.88 bit		

Table 1: Datasets

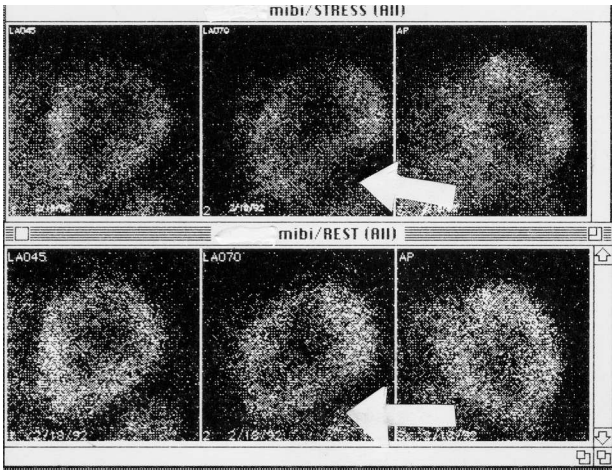

Figure 2: Positive test result – scintigraphic defect seen at stress (upper series) fills at rest (lower series).

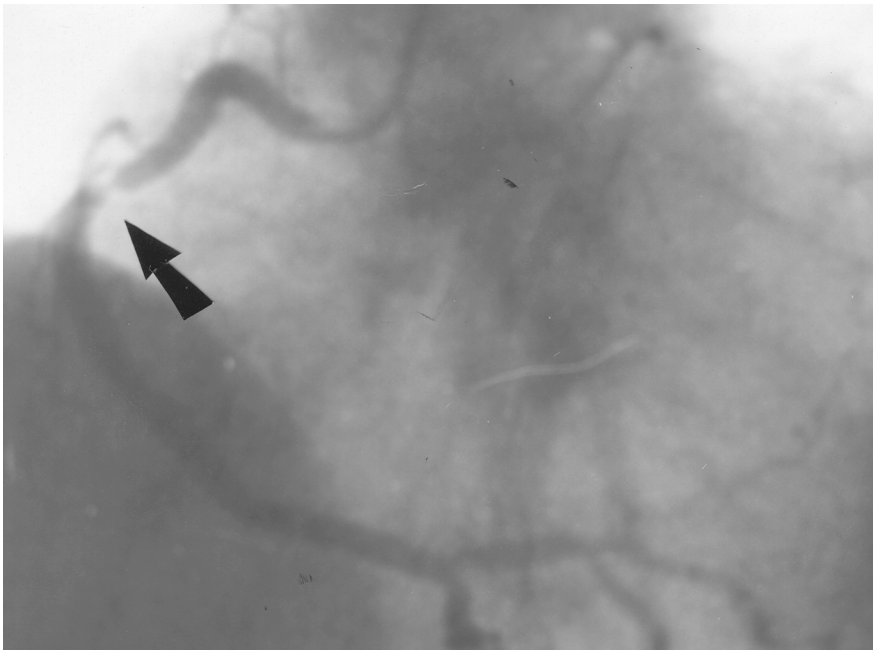

Figure 3: Positive test result. Defect in accumulation of radiopaque materials in right coronary artery caused by arteriosclerotic plaque, causing the stenosis.

3 The algorithms used

In our experiments we used the following algorithms: the naive Bayesian classifier, backpropagation learning of neural networks, two algorithms for induction of decision trees (Assistant-I

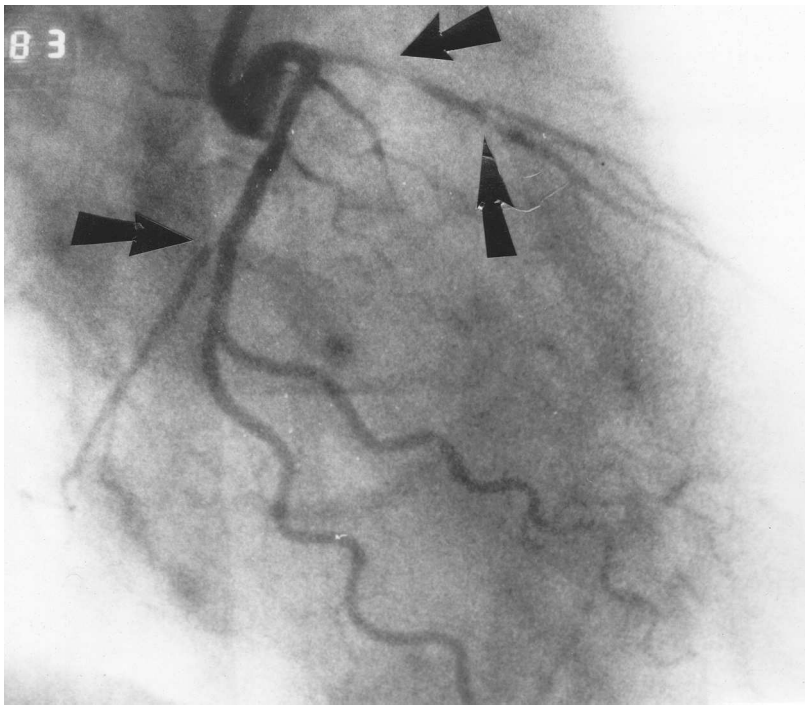

Figure 4: Widespread IHD. Coronary arteries are narrow and tortuous.

and Assistant-R), and *k*-nearest neighbours method.

3.1 Naive and semi-naive Bayesian classifier

The *naive Bayesian classifier* uses the naive Bayes formula (4) to calculate the probability of each class given the values of all the attributes and assuming the conditional independence of the attributes. The attributes are usually defined by a human (especially in medical data), and are therefore relatively independent, as humans tend to think linearly. This is the reason why the naive Bayesian formula (4) often performs well on real-world problems.

$$P(C|A_1..A_{|\mathcal{A}|}) = P(C)\prod_{A\in\mathcal{A}} Q_A \quad \text{where} \quad Q_A = \frac{P(A|C)}{P(A)} = \frac{P(C|A)}{P(C)} \tag{4}$$

A new instance is classified into the class with maximum calculated probability. For estimations of prior probabilities $P(C)$ the Laplace's law of succession was used:

$$P(C) = \frac{N(C)+1}{N+N(C)} \tag{5}$$

For estimations of conditional propabilities $P(C|A)$ the *m*-estimate [Cestnik, 1990] was used:

$$P(C|A) = \frac{N(CA)+m\times P_a(C)}{N(A)+m} = \frac{N(CA)}{N(A)+m} + \frac{m\times P_a(C)}{N(A)+m} \tag{6}$$

The parameter m balances between the contributions of the relative frequency $N(CA)/N(A)$ and the prior probability $P_a(C)$. Both Laplace's law of succession and m-estimate are very useful, especially when estimating probabilities from small datasets. In our experiments, the parameter m was set to 2. This setting is usually used as default and, empirically, gives satisfactory results [Cestnik, 1990] although with tuning better results might be expected.

The *semi-naive Bayesian classifier*, described in more detail in [Kononenko, 1991], attempts to balance between the non-naivety and the reliability of approximations of probabilities. When calculating the probability of class C_j in (4) the influence of attributes A_i and A_l is defined by:

$$\frac{P(C_j|A_i)}{P(C_j)} \times \frac{P(C_j|A_l)}{P(C_j)} \tag{7}$$

If, instead of assuming the independence of values A_i and A_l, the values are interdependent, the corrected influence is given by:

$$\frac{P(C_j|A_iA_l)}{P(C_j)} \tag{8}$$

To combine two values together, the following two conditions should be satisfied:

1. the values of (7) and (8) should be sufficiently different
2. the approximation of $P(C_j|A_iA_l)$ should be sufficiently reliable.

The semi-naive Bayesian classifier uses the Chebishev Theorem to calculate the reliability of probability estimates. The difference between (7) and (8) is used as a parameter in this calculation [Kononenko, 1991].

3.2 Backpropagation learning of neural networks

A multilayered feedforward neural network [Rumelhart and McClelland, 1986] is a hierarchical network consisting of fully interconnected *layers* of processing *units* (often called *neurons*). The output of each unit is connected to every unit in the next layer. A network consists of at least two layers – the input and the output. However, this kind of network is able to solve only a very limited class of problems. For a more general network the number and the size of hidden layers between the input and output layers has to be chosen. Connections between units are often referred to as *synapses*, giving a loose analogy with the brain structure. Each connection and unit has a real-valued weight or bias attached to it.

The backpropagation learning procedure minimizes the squared error accumulated from all training instances by implementing a gradient descent on the error surface. The learning procedure basically feeds the input vector to the network, calculates the output vector of the network and compares it to the desired output vector. Based on the difference, the backpropagation procedure performs a gradient descent in the weight space by modifying the synaptic weights with

the *delta rule* [Rumelhart and McClelland, 1986, Haykin, 1994]. It computes the δ's (i.e., the local gradients) and proceeds backward, layer by layer, starting with the output layer. The most annoying problem of backpropagation is *overfitting* the training data. The trained network may become too specialized for describing training instances, therefore being unable to successfully classify unseen instances. This phenomenon is usually a consequence of using an oversized network with too many hidden units. This problem can be at least partially solved by using the methods for early termination of the training procedure (e.g. observing the overfitting on the validation subset) or by *weight elimination* [Weigand et al., 1990]. We experimented with both approaches, however better results were obtained with the first (the validation subset).

3.3 Assistant-R and Assistant-I

Assistant-R [Kononenko et al., 1997] is a reimplementation of the Assistant learning system for top down induction of decision trees [Cestnik et al., 1987]. The basic algorithm goes back to CLS (Concept Learning System) developed by Hunt [Hunt et al., 1966] and reimplemented by several authors (see [Quinlan, 1986] for an overview). The main features of the original Assistant are binarization of attributes (i.e. grouping available attribute values in two subsets, resulting in binary decision tree), decision tree pruning [Niblett and Bratko, 1986], incomplete data handling and the use of the naive Bayesian classifier when there are some attribute values for which no training instances are available.

The main difference between Assistant and its reimplementation Assistant-R is that instead of the information gain, Relief-F [Kononenko, 1994] is used for attribute selection. Relief-F is an improved and extended version of Relief [Kira and Rendell, 1992]. Its key idea is to estimate attributes according to how well their values distinguish between the instances that are near to each other.

Assistant-R also uses the the *m*-estimate [Cestnik, 1990] for reliable estimation of conditional probabilities during building and pruning of the decision tree. The *m*-estimate is also used in the naive Bayes formula and for postpruning.

Assistant-I is a variant of Assistant-R that, instead of Relief-F, uses the information gain [Quinlan, 1986] for the selection criterion, as does the original Assistant.

3.4 K nearest neighbours

The *k*-nearest neighbours algorithm originates in the field of pattern recognition. For a given new instance this algorithm searches for the *k* nearest training instances and classifies the instance into the most frequent class of these *k* instances. In our experiments, the distance *diff* between instances was a combination of the the Manhattan distance (for nominal attributes) and normalized Euclidean distance (for numerical attributes).

Originally, the algorithm predicts only the class of the new instance. However, it is often useful to estimate the class probability distributions. To achieve this goal, the influences of the nearest neighbours (y_j) are suitably weighted, according to their difference to the new instances (x_i). We used a kernel-type smoother that is, a type of local average smoother that, for each target point x_i in attribute space, calculates a weighted average of the target points y_j in its neighbourhood.

The intuitive sense of the kernel estimate is clear: Values of *diff* such that y_j is close to x_i get relatively heavy weights, while values of *diff* such that y_j is far from x_i get small or zero weight. The kernel parameter *kr* controls the size of the region around x_i for which y_j receives relatively large weights. Since bias increases and variance decreases with increasing kernel *kr*, selection of *kr* is a compromise between bias and variance in order to achieve small mean squared error. In practice this is usually done by trial and error.

A suitable kernel function is the *normal kernel function*:

$$w_i = \frac{1}{\sqrt{2\Pi}kr} \sum_{\text{class}(m)=i} e^{-diff^2(m)/2kr^2} \tag{9}$$

The class distribution is calculated by normalizing the weights as follows:

$$\mathbf{P} = (\frac{w_1}{\sum w_i}, \frac{w_2}{\sum w_i}, \ldots, \frac{w_n}{\sum w_i}) \tag{10}$$

Here n is the number of classes. In all our experiments the parameter k (number of nearest neighbours) was set to 5, and *kr* was set to 2.

4 Cost-sensitive learning

In the field of Machine Learning, there has been some work done concerning cost-sensitive learning, starting with Breiman et al. [Breiman et al., 1984], Knoll et al. [Knoll et al., 1994], Pazzani et al. [Pazzani et al., 1994], Provost [Provost and Danyluk, 1996, Provost and Fawcett, 1997] and Turney [Turney, 1995]. There are several articles in which different techniques are suggested [Turney, 1996].

The Machine Learning algorithms that are used for classification (classifiers) are typically designed to minimize the number of errors (incorrect classifications) made. When misclassification costs vary between classes, this approach is not suitable. In this case the total misclassification cost should be minimized. In our case, the sensitivity and the specificity were much more important (especially specificity) than the classification accuracy. The misclassification costs can be changed in order to bias the algorithms towards higher sensitivity or specificity. So we generalized all the described algorithms to take in account the misclassification costs.

4.1 Definitions

For dealing with misclassification costs we need to define the following terms: a *cost matrix*, a *cost vector*, an *uniform* case, and a metric for evaluation of classifiers' performance in the case of non-uniform costs.

The cost of misclassifying an example is a function of the predicted class and the actual class. This function, $cost(actual\ class, predicted\ class)$ is represented as a cost matrix. This cost matrix is an additional input to the learning procedure and is also used to evaluate the ability of the classifier to reduce misclassification costs.

The *cost matrix* is defined as follows:

- $Cost[i, j] =$ cost of misclassifying an example from "class i" as "class j"
- $Cost[i, i] = 0$ (cost of correct classification).

When all costs are equal, we have the *uniform* cost matrix, where all diagonal elements equal to 0, and all non-diagonal to 1:

$$\forall i, j : Cost[i, j] = \begin{cases} 1, & i \neq j \\ 0, & i = j \end{cases} \tag{11}$$

The *cost vector* represents the expected cost of misclassifying an example that belongs to the i-th class:

$$CostVector[i] = \frac{1}{1 - P(i)} \sum_{j \neq i} P(j) Cost[i, j] \tag{12}$$

where $P(i)$ is an estimate of the prior probability that an example belongs to the i-th class. In the equal-cost case we have the *uniform cost vector*:

$$\forall i : CostVector[i] = 1 \tag{13}$$

The performance criterion is no longer the error rate (or classification accuracy), but the *average cost per example*:

$$Average\ cost = \frac{1}{N} \sum_{i=1}^{N} Cost[\text{actual class}(i), \text{predicted class}(i)] \tag{14}$$

where N is the number of testing examples. The error rate may be viewed as a special case of the average cost, when the uniform cost matrix is used.

$$\begin{aligned} Error\_rate &= \frac{\#\ of\ incorrectly\ classified\ examples}{N} \\ Accuracy &= 1.0 - Error\_rate \end{aligned} \tag{15}$$

As a reference point to which all the results are compared, a simple algorithm that predicts the least expected cost class is usually used [Pazzani et al., 1994]. The least expected cost class is found by minimizing the average cost of guessing class c on the training set of size N_T:

$$\textit{Least expected cost} = \min_{c \in \textit{Class}} \frac{1}{N_T} \sum_{i=1}^{N_T} \textit{Cost}[\text{actual class}(i), c] \tag{16}$$

Note that in the uniform case the least expected cost class is equivalent to the *default* class (the class that most frequently occurs in the training set).

4.2 Non-uniform Misclassification Costs in the Machine Learning algorithms

One possible approach to incorporate the misclassification costs in Machine Learning methods is by altering prior probability estimations [Breiman et al., 1984], either by modifying the probability estimations or by weighted sampling. The basic idea is as follows. Suppose we have a two-class problem with equal probabilities and it is twice as expensive to misclassify a "class 1" example than a "class 2" example. In this case we want an algorithm that misclassifies fewer "class 1" examples. Another way to look at it is that every example in "class 1" counts double when misclassified, so the situation is similar to that if the prior probability of the "class 1" would be twice as large as that of the "class 2". In methods, where probabilities are not explicitly estimated, this approach can be simulated with *weighted sampling*. In the spirit of this approach we developed the modifications of the Machine Learning algorithms. Another important goal was, that in the uniform case the behaviour of the modified algorithm should remain identical to that of the original algorithm.

4.2.1 Naive and semi-naive Bayesian classifier

Naive and semi-naive Bayesian classifier estimate prior and conditional probabilities by using the Laplace's law of succession and m-estimate (6), respectively. Thus the corrected estimations, described in Section 4.2.3 can be used (equations (21) and (23)).

The semi-naive Bayesian classifier also uses the corrected estimations when it calculates the reliability of probability estimates for joint attribute values.

4.2.2 Backpropagation learning of the neural networks

The misclassification costs can be taken into account by changing the error function that is being minimized. Instead of minimizing the squared error, the modified backpropagation learning procedure minimizes misclassification costs. The error function is corrected by introducing the

factor $K[i,j]$, $i =$ desired class, $j =$ actual class:

$$E = \sum_{p \in Examples} \frac{1}{2} \sum_{i \in Output} ((y_i - o_i) \cdot K[class(p), i])^2 \tag{17}$$

The factor $K[i,j]$ should be defined in such a way that the behaviour of the backpropagation algorithm in the uniform case remains the same. Depending on the correct class i, two cases are to be considered:

$$K[i,j] = \begin{cases} CostVector[i], & i = j \quad \text{(expected misclassification cost)} \\ \\ Cost[i,j], & i \neq j \quad \text{(actual misclassification cost)} \end{cases} \tag{18}$$

If we look at the derivation of the backpropagation algorithm, we can see that the $K[i,j]$ behaves as a constant factor in the partial derivatives of the error function [Haykin, 1994]. So the delta rule that takes in account the misclassification cost can be written as follows (c is the desired class of the current training example):

$$\delta_j = \begin{cases} (y_j - o_j) \cdot o_j(1 - o_j) \cdot K^2[c,j], & \text{for output neurons} \\ \\ o_j(1 - o_j) \sum_k \delta_k w_{kj}, & \text{for hidden neurons} \end{cases} \tag{19}$$

To ensure the convergence of the modified backpropagation algorithm, the δ factor for output neurons should be normalized with $\max_{i,j} K[i,j]^2$:

$$\delta' = \frac{\delta}{\max_{i,j} K[i,j]^2} \tag{20}$$

4.2.3 Assistant-R and Assistant-I

Most authors [Knoll et al., 1994] utilize misclassification cost information only when pruning trees and ignore it when growing them. By the generalized *altered priors* approach [Breiman et al., 1984], that is, by changing estimates for prior and conditional probabilities, the misclassification costs are taken into account when growing as well as when pruning trees.

Assistant-I Prior probabilities, estimated either with relative frequency or with Laplace's law of succession, are altered as follows:

$$P'(C_j) = \frac{P(C_j) CostVector[C_j]}{\sum_{i=1}^{N} P(C_i) CostVector[C_i]} \tag{21}$$

Conditional probabilities, estimated with the m-estimate [Cestnik, 1990]:

$$P(C_j|A) = \frac{N(C_j,A) + m \cdot P(C_j)}{N(A) + m} \tag{22}$$

are altered in the same manner:

$$P'(C_j|A) = \frac{N(C_j,A) + m \cdot P'(C_j)}{N(A) + m} \Bigg/ \sum_i \frac{N(C_j,A) + m \cdot P'(C_j)}{N(A) + m} \tag{23}$$

Assistant-R The key idea of the algorithm Relief-F is to estimate attributes according to how well their values distinguish among the instances that are near to each other. Relief-F's estimate $W[A]$ of attribute's quality is an approximation of the following difference of probabilities:

$$\begin{aligned} W[A] &= P(\text{different value of A}|\text{nearest instance from different class}) \\ &- P(\text{different value of A}|\text{nearest instance from same class}) \end{aligned}$$

The values of good attributes should distinguish well between instances from different classes and have similar values for instances from the same class. When misclassification cost are not uniform, good attributes should better distinguish between higher cost instances. So, the nearest instances are weighted as follows:

$$weight(R) = \frac{P'(class(R))}{P(class(R))} = \frac{CostVector[class(R)]}{\sum_i P(C_i) CostVector[C_i]} \tag{24}$$

The calculation of the attribute's quality is only slightly changed. In the uniform case the $weight(R) = 1$ and the attribute estimation is unchanged.

```
set all weights W[A] := 0.0;
for i := 1 to n  do
    begin
        randomly select an instance R;
        find nearest hit H and nearest miss M;
        for A := 1  to #all_attributes do
            W[A] := W[A] - weight(R) * (diff(A,R,H)/n + diff(A,R,M)/n);
                          ^^^^^^^^

    end;
```

4.2.4 K nearest neighbours

The classification procedure of the *k*-nearest neighbours algorithm basically consists of summing up the influences of the nearest neighbours. The influence of an example is proportional to its distance to the new instance. Since the correct class is known for all nearest neighbours, their influence can additionally be weighted with the expected misclassification cost.

$$w'_i = \frac{1}{\sqrt{2\Pi}kr} \sum_{\text{class}(m)=i} CostVector[i] \cdot e^{-diff^2(m)/2kr^2} \tag{25}$$

Finally, the probability that the new instance belongs to any class C_i is calculated as

$$P'(C_i) = w'_i \Big/ \sum_j w'_j \tag{26}$$

	Clinicians			
	Accuracy	Inf. score	Specificity	Sensitivity
Exercise ECG only	0.65	0.10	0.76	0.61
Myocardial scintigraphy only	0.83	0.51	0.85	0.83

Table 2: Results obtained by clinicians on our dataset.

5 Experimental results

The learning task for the Machine Learning algorithms was divided into four steps, differing by the amount of clinical and laboratory data available for each patient (see Table 1):

1. Signs and symptoms;
2. Signs, symptoms, and exercise ECG;
3. Signs, symptoms, exercise ECG, and myocardial scintigraphy;
4. Myocardial scintigraphy only.

In the first two cases we compared our results with results obtained by the clinicians from the exercise ECG only. The third and fourth case were compared with the clinicians' results from the myocardial scintigraphy only.

The experiments on each variation of our dataset were performed with 10-fold stratified cross-validation and the results were averaged. Each system used the same training and testing subsets in order to provide the same experimental conditions.

Besides the classification accuracy, sensitivity, and specificity, we measured also the average information score [Kononenko and Bratko, 1991]. This measure eliminates the influence of prior probabilities and appropriately deals with the probabilistic answers of the classifier. The average information score is defined as:

$$Inf = \frac{\sum_{i=1}^{\#testing\ instances} Inf_i}{\#testing\ instances} \tag{27}$$

where the information score of the classification of i-th testing instance is defined by:

$$Inf_i = \begin{cases} -\log_2 P(Cl_i) + \log_2 P'(Cl_i), & P'(Cl_i) \geq P(Cl_i) \\ -(-\log_2(1-P(Cl_i)) + \log_2(1-P'(Cl_i))), & P'(Cl_i) < P(Cl_i) \end{cases} \tag{28}$$

Cl_i is the correct class of the i-th testing instance, $P(Cl)$ is the prior probability of class Cl and $P'(Cl)$ the probability provided by the classifier.

In Tables 2 – 4 the results of clinicians and Machine Learning algorithms are presented and compared. The information score is given in bits, while the other three measures are given as probabilities.

	Accuracy	Inf. score	Specificity	Sensitivity
	Exercise ECG only			
Clinicians	0.65	0.10	0.76	0.61
	Signs and symptoms only			
Backpropagation	0.80	0.59	0.76	0.82
Naive Bayes	0.80	0.65	0.53	0.92
Semi-naive Bayes	0.80	0.65	0.54	0.92
	Signs, symptoms, and exercise ECG			
Naive Bayes	0.82	0.69	0.62	0.90
Semi-naive Bayes	0.81	0.69	0.59	0.90

Table 3: Exercise ECG. The best results of Machine Learning algorithms compared with clinicians.

As we can see, all the algorithms significantly outperform clinicians on all diagnostic levels, especially when using all available data. The most significant improvements were reached by using the backpropagation learning of neural networks and semi-naive Bayesian classifier on all available data. Both significantly outperform clinicians in classification accuracy (0.92 and 0.91, respectively, versus 0.83). However, this holds for classification accuracy only. In our case, the goal was to increase the specificity, even if this means a slight decrease in the specificity. From this point of view the clinicians still performed slightly better (0.85 versus 0.84 and 0.81, respectively), although for the price of significantly lower sensitivity(0.83 versus 0.86 in both cases). Since the algorithms are designed to maximize classification accuracy, we cannot directly influence their sensitivity and specificity. However, this can be done via variable misclassification cost (Section 5.1).

The second interesting result is that using machine learning techniques one can merely from the evaluation of signs and symptoms achieve the specificity of 0.76 and the sensitivity of 0.82 (as achieved by the neural networks). This means that with the same specificity, much higher sensitivity can be reached compared to the results the clinicians obtained from the exercise ECG data (specificity 0.76, sensitivity 0.61). Since the results of the Machine Learning algorithms when evaluating signs, symptoms, and exercise ECG together were almost the same, it seems that the exercise ECG test results do not provide much new information.

5.1 ROC Analysis and experiments with misclassification costs

ROC (receiver operating characteristic) graphs have long been used in signal detection theory to depict trade-offs between hit rate (*sensitivity*) and false alarm rate ($1.0 - $*specificity*). ROC analysis has lately been extended for use in visualizing and analysing the behaviour of diag-

	Accuracy	Inf. score	Specificity	Sensitivity
	Myocardial scintigraphy only			
Clinicians	0.83	0.51	0.85	0.83
Backpropagation	0.90	0.73	0.82	0.93
Semi-naive Bayes	0.90	0.87	0.81	0.94
Assistant-I	0.90	0.78	0.79	0.94
	Signs, symptoms, exercise ECG and myocardial scintigraphy			
Backpropagation	0.92	0.75	0.84	0.96
Semi-naive Bayes	0.91	0.88	0.81	0.96

Table 4: Myocardial scintigraphy. The best results of Machine Learning algorithms compared with clinicians

nostic systems, and is used for visualization in medicine [Provost and Fawcett, 1997], where specificity–sensitivity relations are often analysed.

In the usual setting, the Machine Learning algorithms are tuned to maximize classification accuracy. In our case, the sensitivity and specificity were more important. The clinicians especially wanted to see if it is possible to increase the specificity of the diagnostic process without affecting the sensitivity too much (this may lead to the reduction of number of patients that are being unnecessarily submitted to invasive pre-operative examinations).

The idea was to see whether and how much we can influence the behaviour of the algorithms (with cost-sensitive modifications) by changing the misclassification costs in favour of one or another class. By giving examples from the *negative* class higher costs, the classifier should become more specific and less sensitive. On the other hand, by giving examples from the *positive* class higher costs, the classifier should become more sensitive and less specific. Note that the modifications of algorithms (Section 4) are not limited to the two-class problems. However, in the multi-class case the results cannot be presented in the ROC graphs.

In our experiments, the misclassification costs varied between $1:20$ in favour of the *negative* class (no IHD present; higher specificity) and $20:1$ in favour of the *positive* class (IHD present; higher sensitivity). The results of our experiments with different algorithms are presented in Figures 5 – 16 Each algorithm's behaviour is shown in two figures. The first one depicts classification accuracy, sensitivity and specificity. The vertical line marks the *uniform cost* $(1:1)$ situation (behaviour of the unmodified algorithm). The second figure shows the ROC (receiver operating characteristic) curve, that is, a trade-off between sensitivity and specificity. By changing the misclassification costs, one actually traverses along this curve. The results shown are averages of the ten-fold stratified cross validation on the complete dataset (all diagnostic levels,

77 attributes, 327 examples).

From the Figures 5 – 16 it can be seen how even a slight increase of specificity drastically reduces sensitivity and vice versa. The most suitable point on the ROC curve was selected by the clinician (Figure 8). It is not necessarily the point where the highest classification accuracy was reached. This point shows the 0.03 increase in specificity and 0.06 in sensitivity, when compared with clinicians' results on the same data. It was achieved by the naive Bayesian classifier (Figures 7 and 8). The misclassification costs were slightly in favour of the negative class. The achieved classification accuracy was 0.88, specificity 0.88 and sensitivity 0.89.

In the clinicians' opinion these results are good enough to be useful in practice, especially when viewed in the spirit of their predictive power (Section 6).

Another interesting and somewhat surprising result is that sometimes, by slightly changing the misclassification costs in favour of the positive (majority) class, the classification accuracy actually improves, compared to the uniform costs situation. This behaviour is unexpected, since the uniform costs situation corresponds to the unmodified algorithms, as they are designed to maximize the classification accuracy. We also experimented with other domains, but this phenomenon seems to be limited to the IHD dataset, described in this paper.

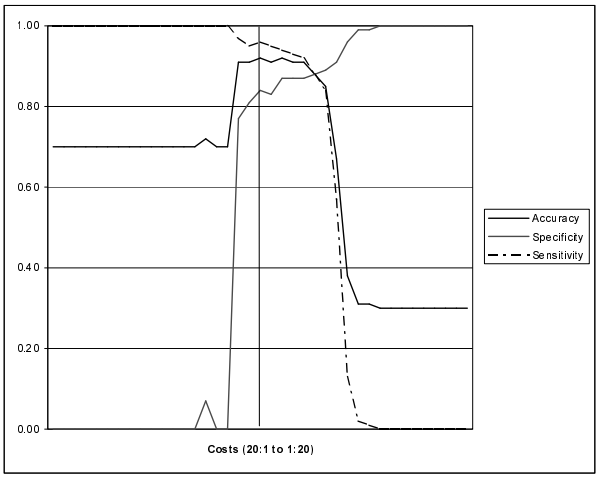

Figure 5: Backpropagation: accuracy, specificity and sensitivity

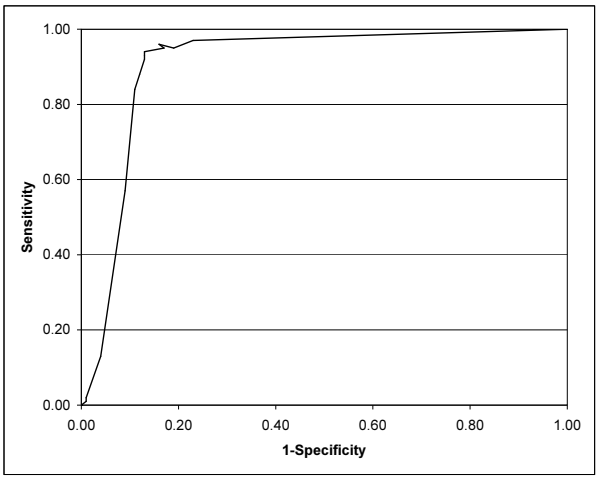

Figure 6: Backpropagation: the ROC graph

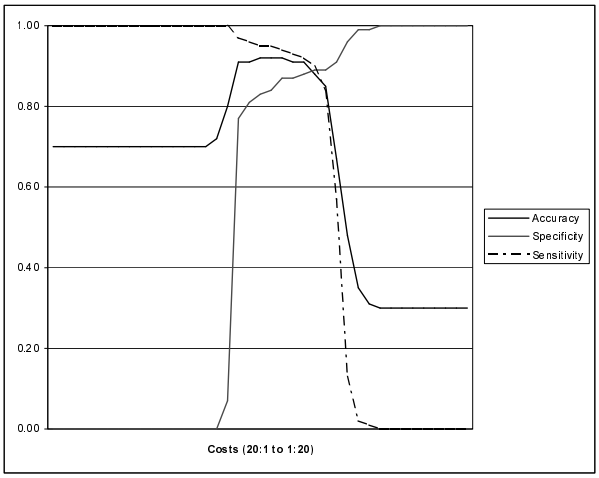

Figure 7: Naive Bayes: accuracy, specificity and sensitivity

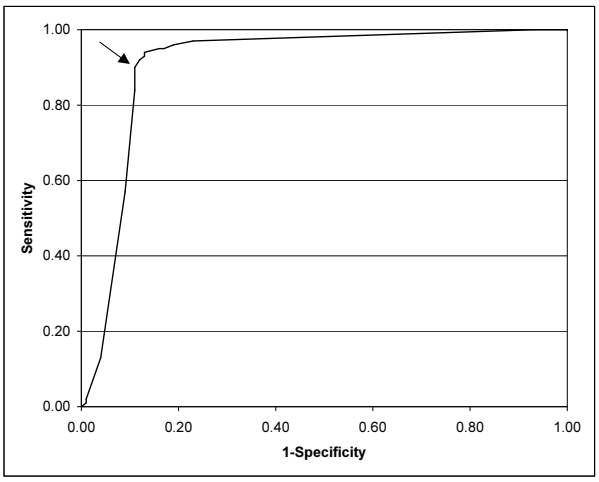

Figure 8: Naive Bayes: the ROC graph. The arrow indicates the result selected by the clinician as best overall.

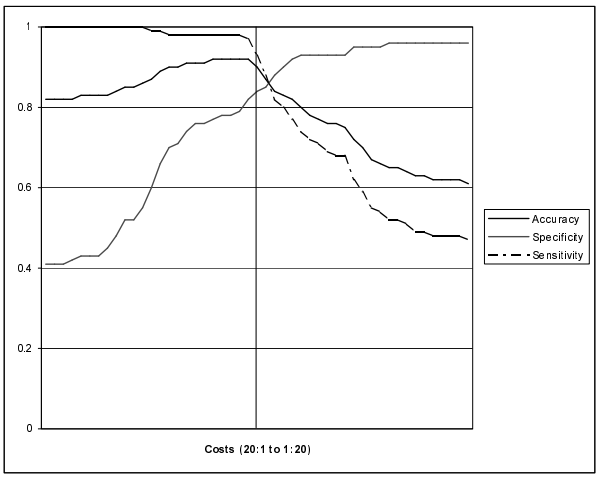

Figure 9: Semi-naive Bayes: accuracy, specificity and sensitivity

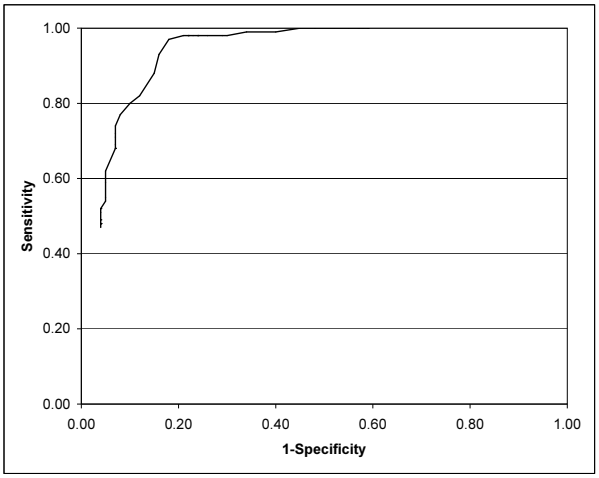

Figure 10: Semi-naive Bayes: the ROC graph

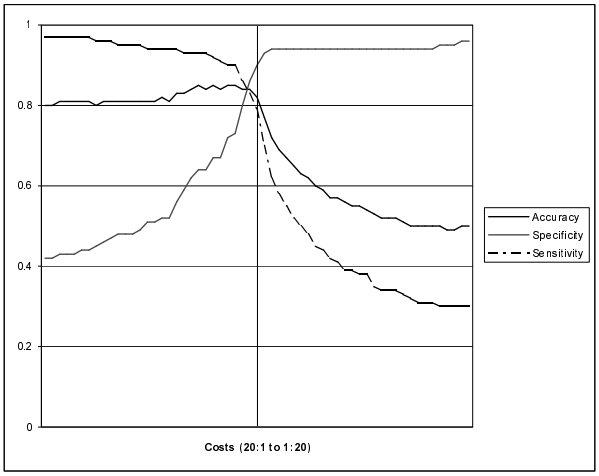

Figure 11: K-nearest (K=5): accuracy, specificity and sensitivity

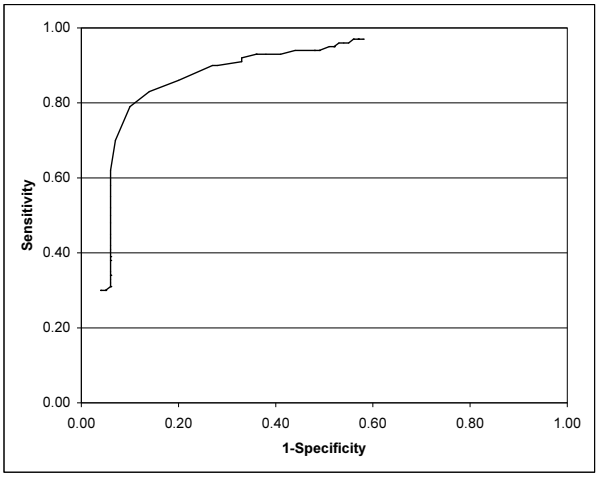

Figure 12: K-nearest (K=5): the ROC graph

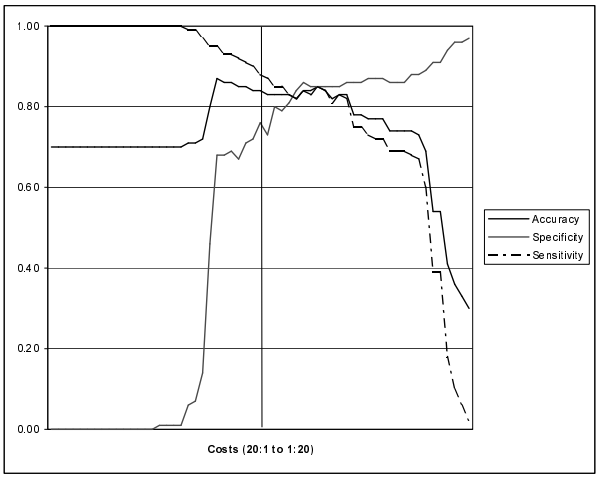

Figure 13: Assistant-R: accuracy, specificity and sensitivity

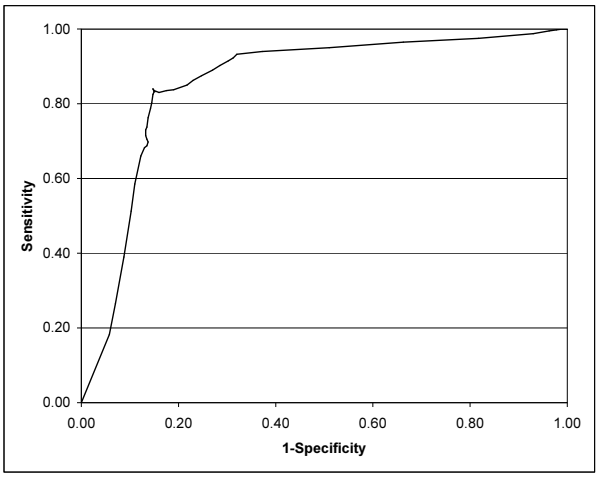

Figure 14: Assistant-R: the ROC graph

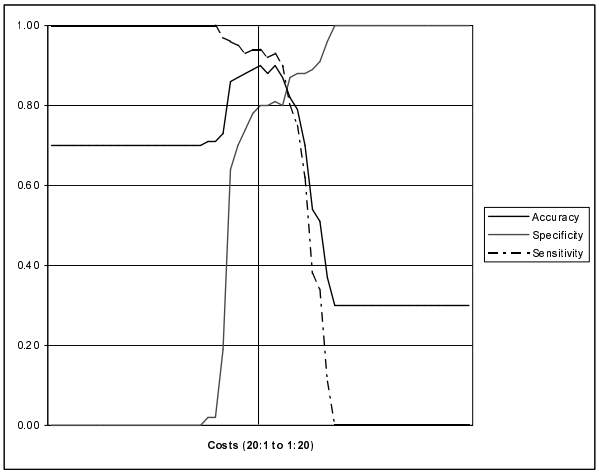

Figure 15: Assistant-I: accuracy, specificity and sensitivity

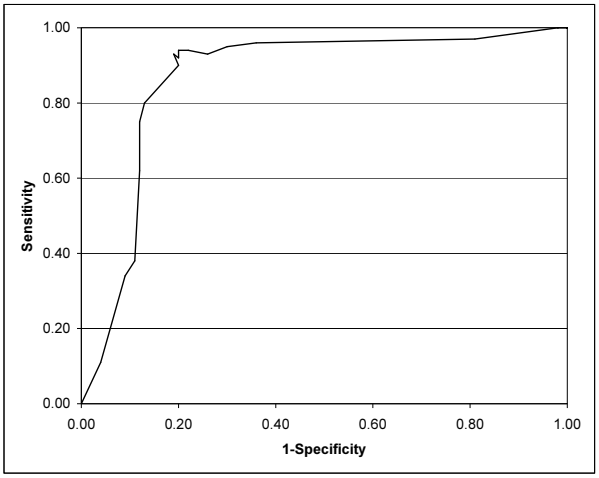

Figure 16: Assistant-I: the ROC graph

	Inf.gain	χ^2	Relief-F
Backpropagation	0.92	0.93	0.92
Naive Bayes	0.92	0.91	0.91
Semi-naive Bayes	0.92	0.92	**0.92**
K-nearest (K=5)	0.90	0.91	0.91
Assistant-R	0.91	0.90	0.90
Assistant-I	0.90	0.91	0.90

Table 5: Maximum accuracy, achieved with different algorithms and estimates. The result with th best accuracy : number of attributes used ratio is emphasised.

5.2 Feature subset selection

Since the amount of data describing each patient is too large for reliable and objective evaluation by clinicians, we wanted to determine the minimum number of attributes, where the accuracy, sensitivity, and specificity reach their maximum. We were interested in reaching as high accuracy, sensitivity and specificity as possible with as few attributes as possible.

We experimented with three different measures for assessing the attribute's importance: information gain [Quinlan, 1986], Relief-F [Kononenko et al., 1997] and χ^2 statistics [Chase and Brown, 1986, pages 589–593]. The results were obtained by the ten-fold stratified cross-validation procedure. All the experiments were performed on the complete dataset (all diagnostic levels, 77 attributes, 327 examples). We ordered the attributes according to the used measure and then trained and tested the classifiers by gradually increasing the number of attributes. The classification accuracy, specificity and sensitivity, achieved with different number of attributes, were measured.

As we can see from Table 5, all the algorithms using different orderings of attributes achieved approximately the same maximum classification accuracy. However, Figure 17 shows that the accuracy of 0.92 can be achieved with as low as 10 attributes (also, the accuracy remains that high up to 31 attributes). Decision tree learners and *K*-nearest neighbours are superior to other algorithms with respect to obtaining maximum accuracy with few attributes. Relief-F seems to be the most appropriate measure for ordering the attributes, because with its ordering, the algorithms achieve the maximum accuracy with (on average) lowest number of attributes.

Overall, the accuracy of 0.92, achieved by the semi-naive Bayes with Relief-F's ordering on only 10 attributes, seems to be the best result. Figure 18 depicts the accuracy of different algorithms when increasing the number of used attributes (Relief-F's ordering).

However, we should not concentrate only on the classification accuracy, because in our case, the specificity and sensitivity are more important. The best result was obtained with the semi-naive Bayesian classifier using Relief-F's ordering of attributes (Figure 18). With only 10

attributes the achieved specificity was 0.84 and sensitivity was 0.96.

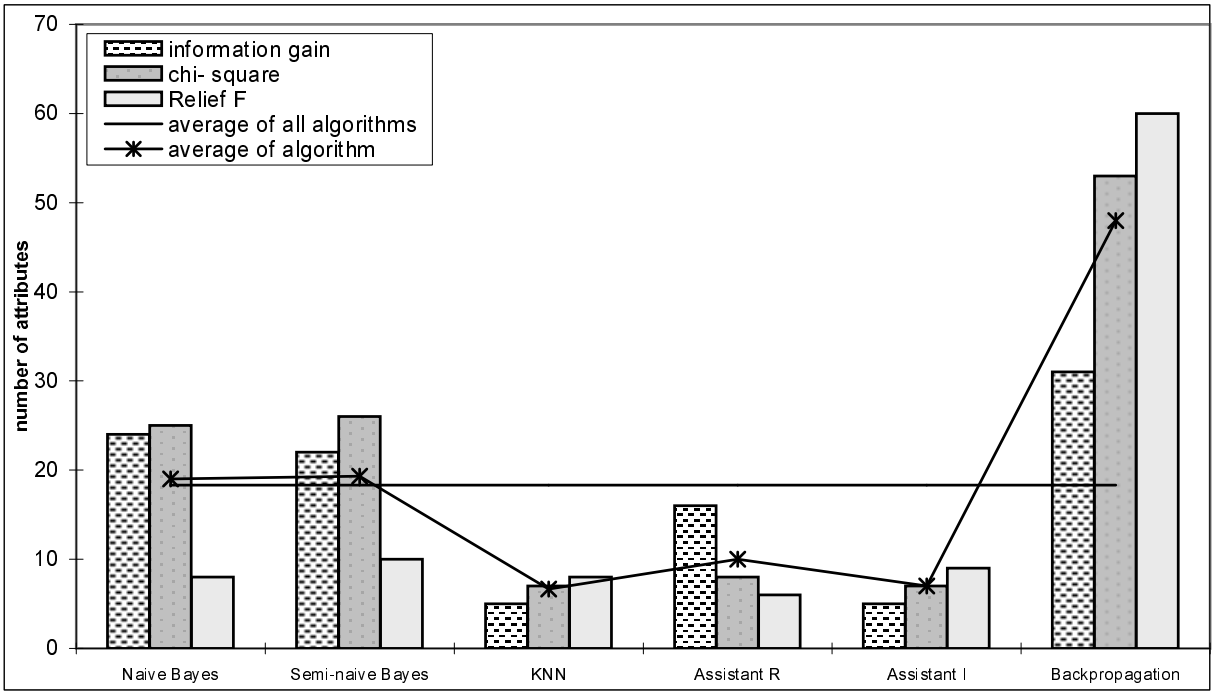

Figure 17: The minimum number of attributes, where maximum accuracy was first achieved, using different algorithms and estimates.

Used attributes come mainly from the third diagnostic level (myocardial scintigraphy): 9 out of 10 best estimated attributes (with Relief-F) come from this level and one comes from the first level (AP – angina pectoris – how typical the chest pain is). If we look at the first 20 attributes, the only attributes that don't come from the third level are:

- from first level AP in the 1st spot, MI(previous myocardial infarction) in the 12th spot, and previous invasive procedures performed on the patient in the 18th spot.
- From the second level(exercise ECG) the ST segment downslopping is in the 14th spot and chest pain during exercise in the 19th spot.

This means, that if we use 10 attributes (the best result), we leave out 28 of 29 attributes from the first diagnostic level, 21 out of 21 attributes from second level and 8 of 17 attributes from third level. When interpreting the results we should keep in mind that the training data includes only those patients with completed all four diagnostic levels.

Thus, many patients, those conclusively diagnosed at the first or second level, are excluded from this study, and our findings may not hold in general. In particular, the first or second diagnostic level cannot be considered unnecessary, although it may appear so from the results.

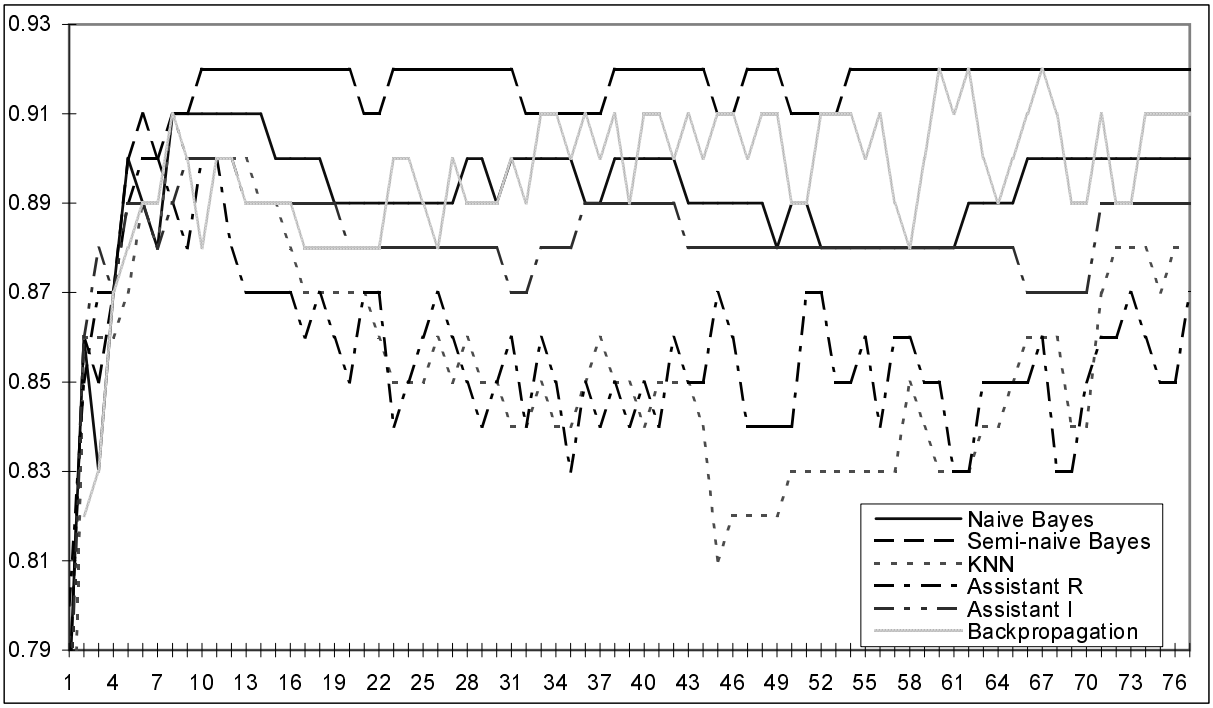

Figure 18: Accuracy, achieved with 6 algorithms using estimate Relief F.

6 Improving the predictive power of tests

Unless we perform the morphological examination such as the coronary angiography, which is 100% sensitive and 100% specific, the test results are not totally reliable and should therefore be interpreted in a probabilistic sense. That means that after a test the probability of the presence of the disease is reported. The post-test probabilities merely indicate the degree of certainty with which the diagnoses are made. The concept of the post-test diagnostic probabilities was first used by Diamond and Forester [Diamond and Forester, 1979]. It states that the predictive value of any diagnostic test is influenced by the prevalence or the pre-test probability (P_1) of the disease among the tested population, by the results of the diagnostic test and by the sensitivity (Se) and the specificity (Sp) of the test. It is easy to calculate the post-test probability (P_2) using the following formulae:

Result of the test	Post − test probability	
Positive	$P_2 = (P_1 * Se)/(P_1 * Se + (1-P_1)(1-Sp))$	(29)
Negative	$P_2 = (P_1 * (1-Se))/(P_1 * (1-Se) + (1-P_1) * Sp)$	

The post-test probability of any patient after a diagnostic test represents the pre-test probability for the subsequent test. This approach in the diagnosis of IHD has the advantage of incorporating not only one or several of the test results but also the data from the patient's history [Diamond and Forester, 1979]. As already stated, the four diagnostic levels for IHD are: signs

	Specificity	Sensitivity	Post-test probability	
			negative	positive
Clinicians	0.85	0.83	0.43	0.75
Naive Bayes	0.88	0.89	0.25	0.90

Table 6: Comparison of average diagnostic value between clinicians and the naive Bayesian classifier.

and symptoms, exercise ECG, myocardial scintigraphy, and coronary angiography. The pre-test probability of IHD is assessed from three variables, age, sex and type of chest pain. It is based on the results obtained from a medical dataset of 4952 patients with angiographically proven IHD [Gerson, 1987]. This value is the pre-test probability for the exercise ECG. The post-test probability of the exercise ECG represents the pre-test probability for the myocardial scintigraphy.

Using this approach the clinician has to decide the level of certainty that he or she requires. It is considered [Diamond et al., 1983] that sufficient diagnostic certainty is reached when the post-test likelihood of IHD is greater than 0.90, or less than 0.10. In the interval between 0.10 and 0.90, the test results are considered as unreliable and further invasive testing is necessary.

6.1 Improving the predictive power with Machine Learning

Our goal was to predict the results of the coronary angiography from all the available data (signs, symptoms, and results of earlier tests - exercise ECG, myocardial scintigraphy) with Machine Learning methods. For comparison with the classical approach we selected the naive Bayesian classifier at its most suitable point of the ROC curve (Figure 8). The results were compared with that of myocardial scintigraphy as the highest step in the classical diagnostic procedure. All available data was used, because Machine Learning methods are not prone to suggestibility. Our hypothesis was that for some unreliable test results the post-test probability would change towards 0.90 (for positive test results) or towards 0.10 (for negative test results).

6.2 Comparison with the classical approach

As we see in the Table 6 the clinicians' sensitivity of myocardial scintigraphy was 0.83 and specificity 0.85. The post-test probability for IHD was 0.75 for positive results and 0.43 for negative ones. With the application of cost-sensitive naive Bayesian classifier [Grošelj et al., 1997] we achieved sensitivity 0.89, specificity 0.88 and the average post-test probability 0.90 for positive and 0.25 for negative results.

The results of our work are promising. In our group of patients the naive Bayesian classifier

	Pre-test (signs and symptoms)	Post-test exercise ECG	Post-test myo-cardial scintigraphy	Post-test naive Bayes
Diagnosis	positive	negative	positive	positive
Probability	0.50	0.31	0.73	0.95

Table 7: An example of pre-test and post-test probabilities, together with the results of naive Bayesian classifier. Results of coronary angiography were: stenosis LAD $>$ 75%, stenosis LCX $>$ 75%, stenosis RCA$<$ 50%. Conclusive diagnosis: IHD present.

significantly improved the diagnostic accuracy of the myocardial scintigraphy. When compared to the standard diagnostic approach, the naive Bayesian classifier shows significant improvements in sensitivity ($0.83 \rightarrow 0.89$) as well as in specificity ($0.85 \rightarrow 0.88$). These results are even more promising when we observe them in the sense of post-test probabilities. For our group of patients the average positive probability for IHD increased from 0.75 to 0.90, and negative decreased from 0.43 to 0.25.

What does this mean? As already mentioned, the non-invasively obtained diagnosis always deals with probability of persistence of the disease. A post-test probability greater than 0.90 is sufficient to confirms the presence of the disease and a post-test probability under 0.10 its absence. Further diagnostic procedures are not necessary in these situations. The only additional diagnostic procedure indicated is eventual intervention therapy. In our case, in 34 patients (10.4%) the post-test probability changed to a value of over 0.90, in 6 patients (1.8%) from over 0.90 to under 0.90, in 17 patients (5.2%) to under 0.10 and in 5 patients (1.5%) from under 0.10 to over 0.10. This means that potentially 12.2% fewer patients would have to be examined with other (invasive) tests. In this way, a significant impact on the accuracy and rationalisation of the diagnosis of IHD can be achieved by using Machine Learning methods.

7 Discussion

The results of our study are promising. The increase of specificity and sensitivity of the myocardial scintigraphy by using the information from the evaluation of signs and symptoms and from the exercise ECG, is a significant result. The naive Bayesian classifier increased the specificity by 0.03 and the sensitivity by 0.06. If such a system were implemented in practice two-fold rationalization might be expected. Due to higher specificity fewer patients without the disease would have to be examined with coronary angiography which is invasive and therefore dangerous. Together with higher sensitivity this would also save money and shorten the waiting times of the truly ill patients

The second interesting result is that using machine learning techniques one can merely from

the evaluation of signs and symptoms achieve the classification accuracy of 0.80, specificity of 0.76 and sensitivity of 0.82 (as achieved by the backpropagation learning of neural networks). This is, because of higher sensitivity (by 0.21 higher), a much better result than that of clinicians when evaluating the exercise ECG. The fact that the exercise ECG does not provide much new information is known to clinicians, but it holds only for highly experienced specialists. Less experienced medical doctors need the exercise ECG result for reliable diagnostics. By using Machine Learning techniques this (time consuming) test may be avoided.

The third interesting result is that, with only 10 attributes, the maximum accuracy of the test can be reached. A closer look at the structure of this subset of attributes suggests that most of the original 77 attributes are redundant in the diagnostic process. Although only a handful of attributes from the lower diagnostic levels are considered as important (e.g., Angina Pectoris and ST segment downslopping), they significantly contribute to the improved diagnostic performance of the test and therefore shouldn't be excluded from the diagnostic process.

The most significant result of our study are the improvements in the predictive power of the diagnostic process. The 12.2% of the patients who would not need to be examined with costly further tests, represents a significant improvement in the diagnostic power as well as in the rationalization of the existing IHD diagnostic procedure.

However, it should be emphasised that the results of our study are obtained on a significantly restricted population and therefore may not be generally applicable to the normal population, i.e. the patients coming to the Nuclear Medicine Department. Further studies might be needed to verify our findings. In particular, on-line data gathering is necessary to obtain a representative dataset.

Acknowledgements

We thank the reviewers for insightful comments which significantly helped to improve the paper. This work was supported by the Slovenian Ministry of Science and Technology.

References

L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth International Group, Belmont CA, 1984.

B. Cestnik. Estimating probabilities: A crucial task in machine learning. In *Proc. European Conference on Artificial Intelligence 1990*, pages 147–149, Stockholm, Sweden, 1990.

B. Cestnik, I. Kononenko, and I. Bratko. ASSISTANT 86: A knowledge elicitation tool for sophisticated users. In I. Bratko and N. Lavrač, editors, *Progress in Machine Learning*. Sigma Press, Wilmslow, England, 1987.

W. Chase and F. Brown. *General Statistics: Second Edition*. John Wiley & Sons, 1986.

G. A. Diamond and J. S. Forester. Analysis of probability as an aid in the clinical diagnosis of coronary artery disease. *New England Journal of Medicine*, 300:1350, 1979.

G. A. Diamond, H. M. Staniloff, and J. S. Forester. Computer-assisted diagnosis in the non-invasive evaluation of patients with suspected coronary artery disease. *Journal of Am. Cardiol.*, 444, 1983.

C. M. Gerson. Test accuracy, test selection, and test result interpretation in chronic coronary artery disease. In C. M. Gerson, editor, *Cardiac Nuclear Medicine*, pages 309–347. Mc Graw Hill, New York, 1987.

C. Grošelj, M. Kukar, J. Fettich, and I. Kononenko. Machine learning improves the accuracy of coronary artery disease diagnostic methods. In *Proc. Computers in Cardiology*, volume 24, pages 57–60, Lund, Sweden, 1997.

S. Haykin. *Neural Networks: A Comprehensive Foundation.* Macmillan College Publishing Company, New York, 1994.

E. Hunt, J. Martin, and P. Stone. *Experiments in Induction.* Academic Press, New York, 1966.

K. Kira and L. Rendell. A practical approach to feature selection. In D. Sleeman and P. Edwards, editors, *Proc. Intern. Conf. on Machine Learning*, pages 249–256, Aberdeen, UK, 1992. Morgan Kaufmann.

U. Knoll, G. Nakhaeizadeh, and B. Tausend. Cost-sensitive pruning of decision trees. In *Proc. ECML'94*, 1994.

I. Kononenko and I. Bratko. Information based evaluation criterion for classifier's performance. *Machine Learning*, 6:67–80, 1991.

I. Kononenko, E. Šimec, and M. Robnik-Šikonja. Overcoming the myopia of inductive learning algorithms with ReliefF. *Applied Intelligence*, 7:39–55, 1997.

I. Kononenko. Semi-naive Bayesian classifier. In Y. Kodratoff, editor, *Proc. European Working Session on Learning-91*, pages 206–219, Porto, Potrugal, 1991. Springer-Verlag.

I. Kononenko. Estimating attributes: Analysis and extensions of RELIEF. In L. De Raedt and F. Bergadano, editors, *Proc. European Conf. on Machine Learning*, pages 171–182, Catania, Italy, 1994. Springer-Verlag.

M. Kukar, C. Grošelj, I. Kononenko, and J. Fettich. An application of machine learning in the diagnosis of ischaemic heart disease. In *Proc. Sixth European Conference of AI in Medicine Europe (AIME 97)*, Grenoble, France, 1997.

T. Niblett and I. Bratko. Learning decision rules in noisy domains. In *Proc. Expert Systems 86*, Brighton, UK, 1986.

M. Pazzani, C. Merz, P. Murphy, K. Ali ant T. Hume, and C. Brunk. Reducing misclassification costs: Knowledge-intensive approaches to learning from noisy data. In *Proc. 11th International Conference on Machine Learning*, pages 217–225, 1994.

F. J. Provost and A. P. Danyluk. A study of complications in real-world machine learning. `http://www.croftj.net/~fawcett/papers/Provost-Danyluk-96.ps.gz`, 1996.

F. J. Provost and T. Fawcett. Analysis and visualization of classifier performance: Comparision under imprecise class and cost distributions. In *Proc 3rd International Conference on Knowledge Discovery and Data Mining (KDD-97).* AAAI Press, 1997.

J.R. Quinlan. Induction of decision trees. *Machine Learning*, 1:81–106, 1986.

D.E. Rumelhart and J. L. McClelland. *Parallel Distributed Processing*, volume 1: Foundations. MIT Press, Cambridge, 1986.

P. D. Turney. Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm. *Journal of Artificial Intelligence Research*, 2:369–409, 1995.

P. D. Turney. Cost-sensitive learning bibliography. `http://ai.iit.nrc.ca/bibliographies/cost-sensitive.html`, 1996.

S. Weigand, A. Huberman, and D. E. Rumelhart. Predicting the future: a connectionist approach. *International Journal of Neural Systems*, 1(3), 1990.