An Application of Machine Learning in the Diagnosis of Ischaemic Heart Disease

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

Abstract

Ishaemic 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. The diagnostic process is stepwise and the results are interpreted hierarchically, i.e. the next step is necessary only if the results of the former are inconclusive. Because the suggestibility is possible, the results of each step are interpreted individually and only the results of the highest step are valid. On the other hand, Machine Learning methods may be able of objective interpretation of all available results for the same patient and in this way increase the diagnostic accuracy, sensitivity and specificity of each step. In the usual setting, the Machine Learning algorithms are tuned to maximize classification accuracy. In our case, the sensitivity and specificity were much more important, so we generalized the algorithms to take in account the variable misclassification costs. The costs can be tuned in order to bias the algorithms towards higher sensitivity or specificity. We conducted many experiments with four learning algorithms and different variations of our dataset (327 patients with completed diagnostic procedures). Our results show that improvements using Machine Learning techniques are reasonable and might find good use in practice.

1 Introduction

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

The usual procedure in IHD diagnosis consists of four diagnostic levels containing 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 the suggestibility is possible, the results of each step are interpreted individually and only the results of the highest step are valid. The amount of data available for each patient in all diagnostic levels is too large to be efficiently and objectively evaluated by physicians.

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.

1063-7125/97 $10.00 © 1997

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 different diagnostic methods is usually described as classification accuracy (in Machine Learning) or as sensitivity and specificity (in medicine):

accuracy = (true positive+true negative test results) / all patients
sensitivity = true positive test results / all patients with disease
specificity = true negative test results / all patients without disease

The reported average values of these measures, taken from 29 reports containing several thousands of patients are as follows [4]. Sensitivity for exercise ECG (5796 persons) is 72%, specificity 79%, and accuracy 74%. For exercise myocardial scintigraphy (2413 persons) they are 84%, 88%, and 85%, respectively. Coronary angiography is a reference method.

The goal of our study was to improve the diagnostic performance (sensitivity and specificity) of non-invasive diagnostic methods (especially myocardial scintigraphy) by evaluating all available diagnostic information using Machine Learning (ML) techniques. The results of coronary angiography were used as a definite proof of IHD presence.

2 The Problem and the Dataset

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

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

There are four levels of diagnostics of IHD. First signs and symptoms of the disease are evaluated clinically and ECG is performed at rest. This is followed by the sequential ECG testing during controlled exercises. If this test is not conclusive, or if additional information regarding the perfusion of the myocardium is needed, myocardial scintigrapy is performed. Radioactive material which accumulates in the heart muscle proportionally to its perfusion is injected into the patient and the images (scintigrams) showing perfusion of the heart muscle during exercise and rest are taken. By comparing both sets of images, the presence, localization and distribution of the ishaemic tissue are determined. If an invasive therapy of the disease is contemplated, i.e. coronary artery bypass surgery, the diagnosis has to be concluded by imaging of the coronary vessels (injecting the contrast material into the coronary vessels and imaging their anatomy by x-ray coronary angiography).

In our study we used a dataset of 327 patients with performed clinical and laboratory examinations, exercise ECG, myocardial scintigraphy and coronary angiography. In 229 cases the disease was angiographically confirmed and in 98 cases it was excluded. The patients were selected from the population of approximately 4000 patients who were examined at the Nuclear Medicine Department in years 1991-1994. For the sake of our study we selected only the patients with complete diagnostic procedures (all four levels). Our experiments were conducted on four problems, depending on the amount of clinical and laboratory data (attributes) available for learning (between 30 and 77 attributes).

3 Algorithms

In our experiments we used the following algorithms: the naive Bayesian classifier [6], backpropagation learning of neural nets [8] with weight elimination [9] and algorithms for induction of decision tree Assistant-I and Assistant-R [5].

3.1 The naive Bayesian classifier

The naive Bayesian classifier uses the naive Bayes formula [6] 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 (often in medicine), and are therefore relatively independent, as humans tend to think linearly. This is the reason why the naive Bayesian formula often performs well on real-world problems. For estimating conditional probabilities the m-estimate [2] was used and the parameter m was set to 2 in all the experiments.

3.2 Backpropagation with weight elimination

A multilayered feedforward neural network [8] is a hierarchical network consisting of fully interconnected *layers* of processing *units* (often called *neurons*). The backpropagation learning procedure minimizes the squared error accumulated from all training instances by implementing a gradient descent on the error surface. Perhaps 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. The *Weight elimination* [9] at least partially overcomes this problem. With a slight change in the error function, the network is forced to keep the weights as small as possible.

3.3 Assistant-R and Assistant-I

Assistant-R [5] is a reimplementation of the Assistant learning system for top down induction of decision trees [3]. The main features of the original Assistant are binarization of attributes, decision tree pruning, incomplete data handling and the use of the naive Bayesian classifier when there are some attribute's 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, ReliefF [7] is used for attribute selection. Its key idea is to estimate attributes according to how well their values distinguish among the instances that are near to each other. Assistant-R also uses the the m-estimate [2] for reliable estimation of the probabilities during building and pruning of the decision tree: In our experiments, the parameter m was set to 2. This setting is usually used as default and, empirically, gives satisfactory results [2]. Assistant-I is a variant of Assistant-R that, instead of ReliefF, uses information gain for the selection criterion, as does the original Assistant. However, the other differences remain (i.e. m-estimate of probabilities).

4 Experiments

The learning task for ML algorithms was given as four problems, differing in amount of clinical and laboratory data available for each patient.

1. Signs and symptoms
2. Signs, symptoms and the exercise ECG
3. Signs, symptoms and the exercise ECG and scintigraphy
4. The exercise myocardial scintigraphy only

In the first two cases we compared our results with results obtained by the physicians from the exercise ECG only. The third and fourth step were compared with their results from the myocardial scintigraphy only. The experiments on each variation of our dataset was performed by the 10-fold cross validation and the results were averaged. Each system used the same subsets of instances for learning and for testing in order to provide the same experimental conditions. In Tables 1 and 2 the results of physicians and ML algorithms are presented and compared.

Table 1. Results obtained by physicians verified with coronary angiography

Physicians	Accuracy	Sensit.	Specif.
Exercise ECG only	65.1	89.3	57.1
Exercise myocardial scintigraphy only	83.8	83.7	85.7

Table 2. Results of Machine Learning algorithms. 1 = Signs and symptoms, 2 = Signs, symptoms, and the exercise ECG, 3 = Signs, symptoms, the exercise ECG, and scintigraphy, 4 = The exercise myocardial scintigraphy only.

	Naive Bayes			Neural net		
	Acc.	Sensit.	Specif.	Acc.	Sensit.	Specif.
1.	79.1	89.2	54.5	79.2	85.5	63.8
2.	79.7	89.3	57.1	81.6	88.5	65.0
3.	88.5	91.7	80.1	89.7	93.8	79.5
4.	87.3	90.2	80.1	88.4	92.6	79.1
	Assistant-I			Assistant-R		
	Acc.	Sensit.	Specif.	Acc.	Sensit.	Specif.
1.	71.2	73.4	59.3	73.2	76.1	61.9
2.	70.5	73.2	59.3	73.1	76.8	61.0
3.	89.0	89.1	88.1	86.6	89.6	79.7
4.	87.2	88.9	83.2	84.0	87.4	73.5

5 Incorporating the cost metrics

In the usual setting, the Machine Learning algorithms are tuned to maximize classification accuracy. In our case, the sensitivity and specificity were much more important (especially specificity), so we generalized the algorithms to take in account the misclassification costs. The costs can be tuned in order to bias the algorithms towards higher sensitivity or specificity.

One possible approach to incorporate misclassification costs in Machine Learning is by altering prior and conditional probabilities [1], 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 lok 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 this sense we can define a matrix of misclassification costs (*cost matrix*) as follows:

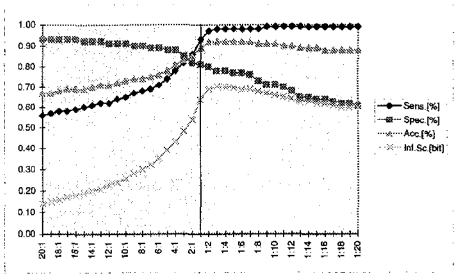
- $Cost[i, j]$ = cost of misclassifying a "class i" example as "class j"
- $Cost[i, i] = 0$ (cost of correct classification).

Bayesian classifier and Assistant-I use the modified probability estimations (Laplace's law of succesion for prior probabilities and m-estimate for conditional probabilities). On the other hand, Assistant-R (ReliefF heuristic) and backpropagation neural network utilize the weighted sampling.

5.1 Experiments with misclassification costs

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 some of the utilized algorithms are shown in Figures 1 - 4. Each algorithm's behaviour is shown in two figures. The first one depicts classification accuracy, information score, sensitivity and specificity. The vertical line marks the *uniform cost* (1 : 1) situation (behaviour of the unmodified algorithm). The second figure shows the ROC (relative operating characteristic) curve, that is, a ratio between sensitivity and specificity. By changing the misclassification costs, one actually traverses along this curve. The results shown are averages of the ten-fold cross validation.



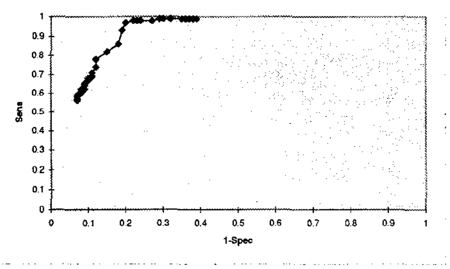
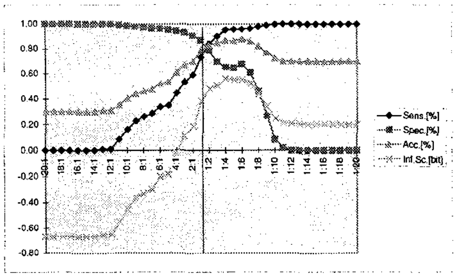


Figure 1. Naive Bayesian classifier

Figure 2. Naive Bayesian classifier - ROC curve



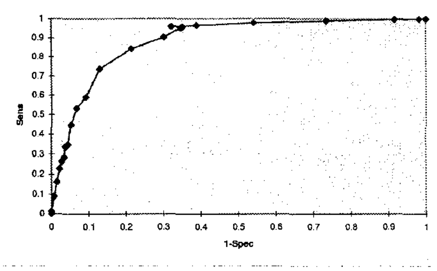


Figure 3. Neural net

Figure 4. Neural net - ROC curve

6 Discussion

The results of our work are promising. The most significant result is the increase of specificity and sensitivity of the exercise myocardial scintigraphy by using other available information (signs, symptoms and exercise ECG). When compared with physicians' results of myocardial scintigraphy, Assistant-I showed the 2.5% increase in specificity and 5.5% in sensitivity. In practice two-fold rationalization could be expected. Due to higher specificity less persons without the disease would have to be examined with invasive and dangerous coronary angiography. Together with higher sensitivity this would also save money and shorten the waiting times of the sick patients The second interesting result is that by using ML techniques one can merely from the evaluation of signs and symptoms achieve the sensitivity of 89% and the specificity of 55% (Bayesian classifier) which is equivalent to the sensitivity and the specificity of the exercise ECG. This fact is well-known but it holds only for experienced physicians specialists. Less experienced physicians need the evaluation of the exercise ECG for reliable diagnostics. By using the ML techniques this could be avoided.

By using the evaluation of the exercise ECG together with the evaluation of the signs and symptoms the neural network increased the specificity for 8% while keeping the sensitivity on the same level (89%). This in turn implies that, if such system was implemented in practice, less persons without the disease would have to pass the myocardial scintigraphy or the coronary angiography.

The possibility of setting variable misclassification costs attracted physicians' attention. In our case they tried to increase the specificity of the diagnostic procedure while trying not to decrease the sensitivity too much. Usually it is not possible to prevent one measure from decreasing when increasing the other one. So by specifying missclassification costs a user actually selects a sensitivity-specificity point on the trade-off curve that fits his/her requirements. This improves the usability of the systems for the medical diagnostic tasks and all problems with variable misclassification costs.

The experiments with variable misclassification costs show that it is possible to increase the classification accuracy compared to the physicians. However, it seems that it is not possible to significantly increase both specificity and sensitivity. In our case only a slight change in cost matrix significantly increased the sensitivity. For similar increases of specificity much bigger changes were necessary. This is in accordance with physicians' speculation that in the problem of IHD diagnostics it is easy to increase sensitivity but hard to increase specificity.

Last but not least, it should be taken into account 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 are needed to verify our findings. In particular, the on-line data gathering is necessary to obtain a representative dataset.

References

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

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

[3] 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.

[4] 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.

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

[6] I. Kononenko. Inductive and Bayesian learning in medical diagnosis. *Applied Intelligence*, 7:317–337, 1993.

[7] 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.

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

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