


---

**Monograph:**
Harrison, R.F., Lim, Chee Peng. and Kennedy, R. Lee. (1994) Autonomously Learning Neural Networks for Clinical Decision Support. Research Report. ACSE Research Report 520 . Department of Automatic Control and Systems Engineering

---

# AUTONOMOUSLY LEARNING NEURAL NETWORKS FOR CLINICAL DECISION SUPPORT

Robert F Harrison†, Chee Peng Lim† and R Lee Kennedy‡

†Department of Automatic Control and Systems Engineering, The University of Sheffield and ‡Department of Medicine, The University of Edinburgh

Research Report No 520

## ABSTRACT

The purpose of this contribution is: to motivate the use of artificial neural networks in "intelligent" clinical decision support; to examine the advantages and limitations of two important classes of artificial neural network; to highlight the potential of intelligent decision support in the early diagnosis of heart attack; and to outline results which indicate, in particular, the potential of the fuzzy ARTMAP network in this acute setting. The work to be described demonstrates that this neural network can overcome problems in knowledge acquisition and portability which may open the way to neural-network-based "apprentices" which learn autonomously whilst providing useful decision support.

## INTRODUCTION

Diagnostic problem solving in whatever domain is a prime example of decision making in the face of uncertainty. Frequently many different outcomes may correspond to identical sets of evidence. The converse may also be true, that any given diagnosis may correspond to a number of distinct sets of evidence. In addition, the data—both outcome and evidence—may be imprecise, adding to the overall uncertainty in the reasoning process and making it probabilistic in nature. These factors can often be the cause of poor diagnostic accuracy in humans, and are, in part, responsible for the difficulties often encountered in developing useful and usable decision support tools.

The appeal of computerized decision support systems in clinical practice is an obvious one, offering the prospect of a fusion of multiple, disparate sources of data and expertise. Furthermore, it is to be expected that by the use of pattern recognition techniques or the methods of artificial intelligence, clinicians might be better served by their decision support systems. That this is so in principle is doubtless true, however, widespread adoption of "intelligent" clinical decision support systems has failed to occur. In diagnosis very few of the intelligent decision support tools that have been developed have ever

ventured beyond their site of origination and have thus had little impact on practice. Why then have computer-based techniques failed to emerge from the laboratory?

1. Data collection and integrity: Classical and Bayesian statistical approaches typically require large cohorts of historical data, both evidence (symptoms, signs, measurements, opinions) and (verified) outcome or diagnosis. Large scale collection of accurate and detailed data can often conflict with clinical practice (*eg* in a busy casualty department). High quality data is necessary not only for the development of statistically-based tools but also for the validation of decision support tools regardless of the underlying technology.
2. Establishing rules or knowledge acquisition: Rule-based systems require a corpus of expert rules which, in conjunction with evidence, may be used to infer outcome. The establishment of a rule base is often fraught with difficulty inasmuch as it is the exceptions (which are of prime importance and which may form a significant proportion of the population) which often elude the knowledge engineer.
3. System portability and adaptability: In both of the previous cases we have identified areas which may make it difficult for the system to be operated away from its site of origination. Significant geographical or demographical variations would require further data capture. Temporal variations in the nature of the problem, on "short" time-scales, may prove intractable. The solutions to these problems may also elude the expert systems developer, where extensive additional investment in knowledge engineering may be required to overcome them.
4. Validation, evaluation, usability and resistance: The problems of how to validate a system's performance and to evaluate its impact in live trials must beset the development of any form of decision aid. In addition, any useful system must impinge as little as possible upon normal working practice. Furthermore, there may be inbuilt resistance to the use of computerized decision support systems such as a fear of the erosion of the expert's rôle.

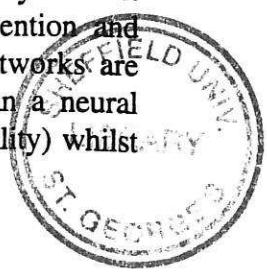
Items 1, 3 and 4 are ever present. Items 2 and 3, however, can be addressed by introducing a "learning" mechanism, which we do here using artificial neural networks.

## NEURAL-NETWORK-BASED DECISION SUPPORT SYSTEMS

Advances in neurocomputing have opened the way for the establishment of decision support systems which are able to learn complex associations by example. The main thrust of work in this area has been in the use of the so-called feedforward networks (*eg* the multilayer Perceptron [1] (MLP) or the Radial Basis Function networks [2] (RBFN)) to learn the association between evidence and outcome. Theoretical work in this area has led to the discovery of two important properties of feedforward networks.

- Their learning rules lead to an interpretation of their outputs as estimates of the posterior probability distribution, conditioned on a set of evidence [3].
- The MLP or the RBFN have been shown to be rich enough in structure so as to be able to approximate any (sufficiently smooth) function with arbitrary accuracy [4,5].

It can be inferred from these results that, given sufficient data, computational resources (the multilayer Perceptron, in particular, does not scale well with problem size) and time (non-linear optimization which is non-linear in the parameters may be time consuming to perform, numerically), it is possible to estimate the Bayes-optimal classifier to any desired degree of accuracy, directly and with no prior assumptions on the probabilistic structure of the data. This is an attractive scenario and has been extensively exploited. However, items 2 and 3 above still remain. The inherent adaptability of NNs may make it easier to tune-in to local conditions but would still require significant intervention and additional effort in data capture, retraining and revalidation. Feedforward networks are static devices in operation and fail to cope with the basic question: "how can a neural network protect from corruption, and retain, useful historical associations (stability) whilst



simultaneously learning new associations (plasticity) which may be unrelated to past experience, or at worst, spurious?". This question is known as the stability/plasticity dilemma. To overcome this problem in feedforward networks, learning is suppressed after acceptable performance is attained. The system is then put into operation. Implicit in this is the assumption that a trained network both represents the problem adequately at the time of development and continues to do so into the future, or in remote situations. Should learning remain continuously active in feedforward networks, all new data will be learned, with the attendant risk of serious performance degradation.

An entirely different approach, utilizing a network comprising both feedforward and feedback components, has been taken by Carpenter and Grossberg and colleagues, and which overcomes to a large extent the stability/plasticity dilemma. This is the so-called Adaptive Resonance Theory (ART) family of architectures [6,7]. In their earliest manifestations these were unsupervised systems which autonomously learned to recognize categories of their own devising. They use feedback to compare the existing state of knowledge or long term memory (LTM) of the system with the current set of evidence and either: (i) adjust the LTM, which codes for a particular category, to account for the current situation if this is "similar" enough to other patterns in that category; or (ii) initiate a new category which codes for the unrecognized (current) pattern. This has a major advantage from a design view point in that there is no off-line "hand crafting" of network architecture to be done, *ie* one autonomous network can address any problem or, indeed, many problems simultaneously. Also, commonly occurring patterns have the effect of reinforcing their category's ability to recognize like examples, while categories representing spurious events are rarely, if ever, excited again and so do not corrupt previously learned information. Conversely, should a rare but valid event occur, it will reside in LTM until next recalled.

The ART architectures of interest here comprise two layers of nodes, fully interconnected in *both* directions. These form an input/comparison field (F1), and an output/recognition field (F2) which latter implements a "winner-take-all" competition. Together these form an *attentional* subsystem which is complemented by an *orienting* subsystem which initiates or suppresses search. ART takes its name from the interplay between learning and recall whereby signals reverberate between the two layers. When an input pattern is recognized, a stable oscillation (resonance) ensues and learning (adaptation) takes place. Categories are coded by the formation of templates in the F2 layer (represented by the weights of a node) and these are refined as new information becomes available. During recall, when a given node is excited, a template is fed-back to the F1 layer for comparison with the current input. The degree of match is assessed against the *vigilance* parameter which is used to control the coarseness of categorization. If the degree of match is not sufficiently good, search is initiated until either an acceptable match is found (resonance) or the pattern is assigned to a new category (F2 node).

Until recently ART was restricted to unsupervised learning. This meant that the autonomously selected categories would be unlikely to correspond to meaningful categories in the problem domain. The so-called ARTMAP [8,9] family of architectures has resolved this problem by providing a mapping network which is capable of supervised learning whilst retaining the desirable properties of the earlier ART networks. These networks comprise two ART modules coupled via a *map field.* Each ART module individually self-organizes into categories representing data (evidence) and supervisory signal (outcome) and the association between categories is formed by the map field.

ARTMAP networks are able to learn to improve their predictive performance on-line in non-stationary environments, utilizing their entire memory capacities. Learning is driven by approximate match and takes place very rapidly as does recall or recognition. Contrast this with the feedforward architectures. These learn off-line and assume a

stationary environment. Learning must be suppressed to overcome the stability/plasticity dilemma and is: very slow, driven by mismatch; prone to spurious solutions; and scales exponentially with problem size. Recall, however, is very fast.

ARTMAP presents the prospect of an autonomous system capable of learning stably to categorize data whilst protecting the user from spurious predictions. This means that the system can safely carry on learning in situ whilst providing useful support. Thus, in clinical diagnosis, evidence would be presented. Should it excite a recognition category (from previous training) then a prediction is returned. Update of LTM can then be initiated if and when diagnosis is confirmed. If the current pattern is not recognized the user is so informed. Again adjustment of LTM is only initiated upon confirmation of the diagnosis. Provided diagnosis remains unconfirmed, no LTM adjustment takes place. This is a crucial issue in the development of a portable decision aid which should be able to adapt to local practice and to changing procedures, in much the same way as humans do.

Any decision making or diagnostic procedure where evidence is to be associated either with an objective outcome or with expert (subjective) opinion, is a potential application area for this approach and most importantly, it can put development of decision aids into the hands of the domain expert, rather than the computing expert. This capability must be viewed as essential in overcoming resistance to the use of computational decision aids—the domain expert assumes "ownership".

We now demonstrate the applicability of an ARTMAP variant, fuzzy ARTMAP [9] (FAM) to the problem of the early diagnosis of myocardial infarction (MI or heart attack). FAM possesses the attractive properties described earlier and overcomes many of the failings of early ART and ARTMAP implementations. FAM achieves a synthesis of fuzzy logic and ART which enables it to learn and to recognize arbitrary sequences of analogue or binary input pairs, which may represent fuzzy or crisp sets of features.

## EARLY DIAGNOSIS OF ACUTE MYOCARDIAL INFARCTION: A CASE STUDY

Early and accurate diagnosis of chest pain is perhaps the major challenge in present day emergency medicine. Chest pain is the commonest reason for emergency medical referral in the developed world and is a major symptom of the onset of MI. Each year in the United Kingdom over 240,000 cases are confirmed while in the United States 1.5 million patients are admitted to intensive therapy units (ITUs). However, in an audit of the management of acute chest pain in a large Accident and Emergency (A&E) department, 12% of patients were found to have been erroneously discharged while 16% were found to have been inappropriately admitted to ITU [10]. In the United States, approaching half of those patients admitted to ITU may ultimately be found not to have suffered a heart attack [11]. Evidently this is a diagnostic question with significant health risk and resource implications. It has been estimated that an early transfer of patients from ITU to a medical ward may result in a financial saving of 50%. Early discharge of those with non-threatening disease would result in concomitantly higher savings [12].

There is further pressure to improve the early diagnosis of MI—the advent of thrombolytic therapy. This is most beneficial when administered as soon after the onset of symptoms as possible. The expected benefit of thrombolytic therapy is a reduction in the immediate threat to life plus an improved long term prognosis. The first of these benefits has been confirmed [13] but it is too early to make a judgement on the second. Thrombolytic agents must be administered within six hours of an infarction taking place to be of significant benefit [14]. Furthermore, they are expensive and may be dangerous if given inappropriately. The routine diagnosis of MI relies on serial measurements of indicators which may take 24–48 hours to develop diagnostic changes. The potential of a

diagnostic aid during the early stages of MI is therefore clear, particularly for clinicians in non-cardiac specialities such as emergency medicine or in general practice.

Acute chest pain is a suitable domain in which to develop a decision support system because of (i) its high incidence and (ii) the possibility of obtaining a concrete final diagnosis. Our study included 500 consecutive emergency referrals to a large teaching hospital in the UK, with a complaint of chest pain [15,16]. Information was recorded from each patient on a standard proforma, comprising 78 items of demographic, clinical and electrocardiographic data. In addition the admitting clinician was asked to estimate the likelihood of the patient having suffered a heart attack. Each proforma was completed before the results from confirmatory tests were available. Twenty six features were abstracted from each patient record and these were coded in a binary vector excepting real-valued data such as age *etc* which were normalised in the range 0–1 [17]. The final diagnoses were assigned independently and were binary coded.

## Methods

A number of experiments have been conducted to explore the capabilities of FAM in the early diagnosis of MI. Three training methods were investigated as follows.

- Single-epoch (SE) training—each pattern pair is presented once only in the following cycle: present evidence, predict outcome, present verified outcome, update long term memory. Such a strategy can be implemented on-line, in real-time.
- Multi-epoch (ME) training—corresponds to the above but the data are presented as many times as is necessary to ensure that as many as possible are correctly classified. Evidently the computational burden rises here with the amount of historical data and thus, for on-line operation the minimum interval between data presentations must increase.
- Voting—is an inherently off-line technique, requiring multiple network realizations using either SE or ME training, for random orderings of the data. The voting strategy can overcome classification errors associated with the order of presentation of data (a known problem with FAM [9]) by cancelling prediction errors.

For investigations into off-line performance the data were partitioned into a training set and an independent test set, as shown in Table 1, below. Note that the data comprises approximately equal proportions of MI, angina and non-ischaemic heart disease (IHD) sufferers and therefore has an *a priori* bias towards excluding a diagnosis of MI of 2.2:1.

A further 26 records were set aside (16 from the training set and 10 from the testing set) to obtain a comparison with a panel of experts. Of these, 20 were thought to be "difficult" cases whilst six were "text-book" cases, included for calibration. Networks were trained on the training set, using the selected method, and diagnostic performance was calculated from the testing set.

In the assessment of on-line performance, the remaining 474 data were used both to train and to test; statistics being gathered prior to verification of diagnosis an LTM update. Here accuracy (ACC), sensitivity (SENS) and specificity (SPEC) of diagnosis are of interest. For off-line training, confidence intervals are calculated for these quantities according to the method due to Highleyman

| | Training Set | | Testing Set | |
|---|---|---|---|---|
| **Final DX** | **Number** | **Proportion** | **Number** | **Proportion** |
| **MI** | 92 | 0.31 | 62 | 0.31 |
| **Angina** | 114 | 0.38 | 75 | 0.38 |
| **Non-IHD** | 94 | 0.31 | 63 | 0.31 |
| **Totals** | 300 | 1.00 | 200 | 1.00 |

Table 1 Classification of chest pain sufferers.

(reported in [18]). Since this method relies upon the division of data into explicit training and testing sets, it is not appropriate to the on-line situation. We have not yet investigated confidence measures for that case.

It should be noted that, whereas it is usual to select optimal decision thresholds by analysis of the Receiver Operating Characteristic (ROC) curve [19], this technique is not appropriate here owing to the "all or nothing" predictions made by the FAM system. It will be seen that this inability to select optimal thresholds, hence to counteract the effects of *a priori* bias in the data, can result in an imbalance in the values of accuracy, sensitivity and specificity. Recent work by the authors introduce a modified FAM which can achieve, on-line, very close to Bayes optimal classification rates for strongly biased data [20].

## Results and Discussion

We concentrate on those results which best illustrate the effectiveness with which FAM addresses items 2 and 3 in the Introduction.

### Off-line processing

Table 2(a) presents accuracies, sensitivities and specificities for the binary decision, MI or not MI for the single-epoch and multi-epoch training strategies and for their associated voting strategies (with ten voters). The columns headed "Ave" contain the means of ten runs (using randomly ordered data) and their associated standard deviations are shown below "SD". Confidence intervals for ACC, SENS and SPEC are also given.

Voting gives the best overall performance and achieves the levelling of ACC, SENS and SPEC usually associated with optimal threshold selection by ROC analysis.

In table 2(b) the performance of the admitting clinicians at presentation, and the best performance achieved (on a super-set of the same data) by an MLP (with optimal decision threshold) [15] are presented.

Clearly, FAM is able to outperform both the admitting clinicians and the best available MLP results when trained off-line. Note that here the admitting clinicians had all received specific cardiological training and thus might be expected to outperform typical casualty officers.

| | SE | | | ME | | |
|---|---|---|---|---|---|---|
| | Ave | SD | Vote | Ave | SD | Vote |
| ACC (%) | $79(^{+6}_{-6})$ | 5 | $86(^{+3}_{-7})$ | $84(^{+3}_{-7})$ | 3 | $90(^{+3}_{-5})$ |
| SENS (%) | $75(^{+6}_{-7})$ | 7 | $86(^{+3}_{-7})$ | $82(^{+6}_{-6})$ | 3 | $90(^{+3}_{-5})$ |
| SPEC (%) | $81(^{+6}_{-6})$ | 9 | $86(^{+3}_{-7})$ | $85(^{+3}_{-7})$ | 4 | $90(^{+3}_{-5})$ |

Table 2(a) Off-line training performance.

| | Clin | MLP |
|---|---|---|
| ACC (%) | 82 | 90 |
| SENS (%) | 79 | 87 |
| SPEC (%) | 84 | 91 |

Table 2(b) Comparisons.

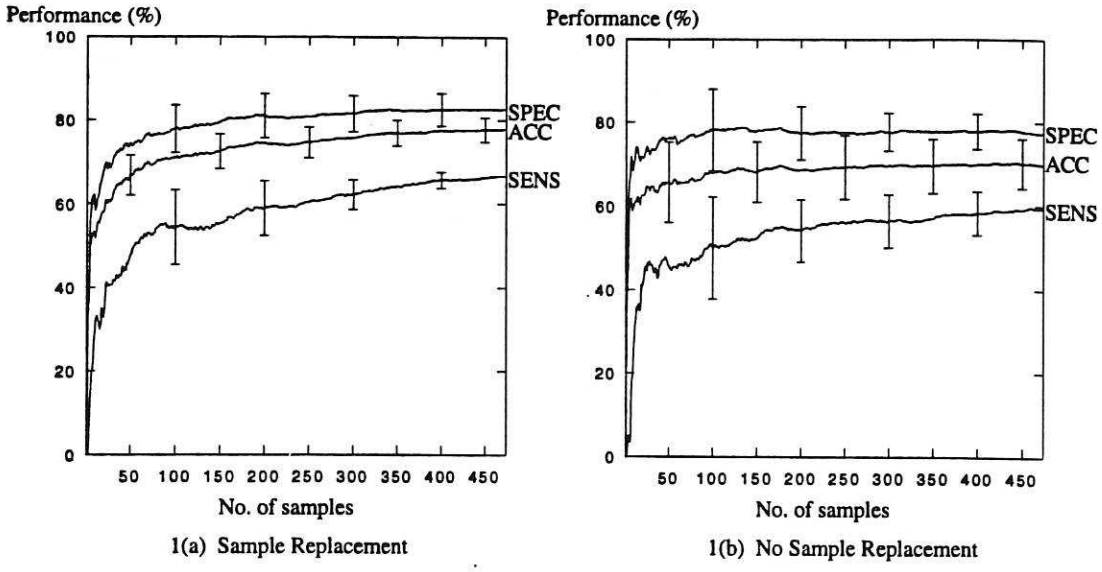
When compared with a panel of experts (senior clinicians with extensive cardiological experience), which was given 26 selected cases to diagnose [15], FAM diagnosed 10/11 cases of MI and excluded 13/15 non-sufferers [17]. The panel diagnosed 10/11 and 9/15 respectively. The relative performance of the panel and FAM is shown in Table 3 which can be analyzed by McNemar's method to produce a chi-square test statistic of 2.66. This is not significant at the 10% level thus we conclude that the panel and FAM perform equally well on the sample.

| | Panel of experts | | |
|---|---|---|---|
| FAM | MI | non-MI | Total |
| MI | 11 | 1 | 12 |
| non-MI | 5 | 9 | 14 |
| Total | 16 | 10 | 26 |

Table 3 DX by FAM and a panel of experts.

### On-line processing

Figure 1 indicates the on-line performance of FAM for two separate cases. The first uses the technique of "sample replacement". Here, samples are drawn at random and are returned after use. Thus any individual sample may be chosen repeatedly. The second case is analogous to in situ or real-time learning when samples are not returned to the pool. Again ACC, SENS and SPEC are considered and their averages over ten runs are plotted with an indication of their standard deviations. There are two important points to note pertaining to on-line processing. (i) Sometimes FAM fails to make a prediction (recognize a pattern). This is especially true in the early stages of learning. Here such non-predictions are counted as erroneous so that the performance indicators are biased downwards. (ii) Because statistics were gathered sequentially for each run, frequent poor (or non)




1(a) Sample Replacement

1(b) No Sample Replacement

Figure 1 On-line FAM performance.

predictions in the early stages are included in the long-run results. Again this has the effect of biasing the results downwards.

In both cases the qualitative behaviour of FAM is as expected: broadly speaking, a continuous improvement in performance as the number of samples increases accompanied by a gradual reduction in spread. This latter indicates that the performance of an individual run has a tendency towards the average performance *ie* in the long run averaging should not be required to achieve good performance and a truly on-line system can be used. Peaks and troughs in the early stages result from the formation of poor initial templates and frequent non-predictions. Sample replacement yields a better result owing to the relatively small sample size (relatively large probability of repetition).

## CONCLUSION

FAM has demonstrated its potential to diagnose acute myocardial infarction. FAM attains the performance of the MLP, and has the following advantages:

- autonomous operation whilst learning to improve predictive performance on-line;
- few parameters to be tuned and little "hand crafting" of architecture;
- recall and learning very rapid (potentially real-time).

The primary disadvantage of FAM *vis a vis* feedforward networks, is its lack of a Bayesian interpretation, which is under further investigation. The authors have shown empirically that FAM can approach the Bayes-optimal solution in on-line mode [20]. It appears, therefore, that FAM may offer solutions to the problems of autonomous machine

acquisition of knowledge and portability, whilst simultaneously providing useful decision support and may thus open the way for a computerized "apprentice" system.

## REFERENCES


1. Rumelhart, D, Hinton, G and Williams, R (1986) Learning representations by back-propagating errors. *Nature*, **323**, 533–536.
2. Moody, J and Darken, C (1989) Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1**, 281–294.
3. Richard, M and Lippman, R (1991) Neural network classifiers estimate Bayesian a posteriori probabilities. *Neural Computation*, **3**, 461–483.
4. Cybenko, G (1989) Approximations by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, **2**, 303–314.
5. Park, J and Sandberg, I (1991) Universal approximation using radial basis function networks. *Neural Computation*, **3**, 246–257.
6. Carpenter, G and Grossberg, S (1987) A massively parallel architecture for a self-organizing neural pattern recognition machine. *Computer Vision, Graphics and Image Processing*, **37**, 54–115.
7. Carpenter, G and Grossberg, S (1987) ART 2: Self-organization of stable category recognition codes for analog input patterns. *Applied Optics*, **26**, 4919–4930.
8. Carpenter, G, Grossberg, S and Reynolds, J (1991) ARTMAP: Supervised real-time learning and classification of nonstationary data by a self-organizing neural network. *Neural Networks*, **4**, 565–588.
9. Carpenter, G, Grossberg, S, Markuzon, N, Reynolds, J and Rosen, D (1992) Fuzzy ARTMAP: A neural network architecture for incremental supervised learning of analog multi-dimensional maps. *IEEE Transactions on Artificial Neural Networks*, **3**, 698–712.
10. Emerson, P, Russell, N, Wyatt, J *et al* (1989) An audit of the management of patients attending an accident and emergency department with chest pain. *Quarterly Journal of Medicine*, **70**, 213–220.
11. Pozen, M, D'Agostino, R, Selker, H *et al* (1984) A predictive instrument to improve coronary care unit admission practices in acute ischaemic heart disease: a prospective multi-centre clinical trial. *New England Journal of Medicine*, **310** 1273–1278.
12. Collinson, P, Ramhamadany, E, Rosalki, S *et al* (1989) Diagnosis of acute myocardial infarction from sequential enzyme measurements obtained within 12 hours of admission to hospital. *Journal of Clinical Pathology*, **42**, 1126–1131.
13. Simoons, M (1989) Thrombolytic therapy in acute myocardial infarction. *Annual Review of Medicine*, **40**, 181–200.
14. McNeill, A, Flannery, D, Wilson, C *et al* (1991) Thrombolytic therapy within one hour of the onset of acute myocardial infarction. *Quarterly Journal of Medicine*, **79**, 487–494.
15. Harrison, R, Marshall, S and Kennedy, R (1991) A connectionist aid to the early diagnosis of myocardial infarction. *Proceedings 3rd European Conference on AI in Medicine*, 119–128.
16. Harrison, R, Marshall, S and Kennedy, R (1994) Neural networks, heart attack and Bayesian decisions: An application of the Boltzmann Perceptron network. *Journal of Artificial Neural Networks*, **1** (in press).
17. Lim, C (1993) An autonomous learning system. MSc dissertation, Department of Automatic Control and Systems Engineering, The University of Sheffield.
18. Duda, R and Hart, P (1973) Pattern classification and scene analysis. Wiley and Sons, New York, 74–75.
19. Meistrell, M (1990) Evaluation of neural network performance by receiver operating characteristic (ROC) analysis: examples from the biotechnology domain. *Computer Methods and Programs in Biomedicine*, 32, 73–80.
20. Lim, C and Harrison, R (1994) Modified fuzzy ARTMAP approaches Bayes optimal classification rates: an empirical demonstration. Research Report 515, Department of Automatic Control and Systems Engineering, The University of Sheffield.