This is a repository copy of *Risk-Sensitive Diagnosis and the Role of Neural Networks*.

White Rose Research Online URL for this paper:
http://eprints.whiterose.ac.uk/79656/

Monograph:

Harrison, R.F., Kennedy, R.Lee. and Marshall, S.J. (1994) Risk-Sensitive Diagnosis and the Role of Neural Networks. Research Report. ACSE Research Report 516 . Department of Automatic Control and Systems Engineering

Reuse

Unless indicated otherwise, fulltext items are protected by copyright with all rights reserved. The copyright exception in section 29 of the Copyright, Designs and Patents Act 1988 allows the making of a single copy solely for the purpose of non-commercial research or private study within the limits of fair dealing. The publisher or other rights-holder may allow further reproduction and re-use of this version - refer to the White Rose Research Online record for this item. Where records identify the publisher as the copyright holder, users can verify any specific terms of use on the publisher’s website.

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

629.8(S)

X

Risk-sensitive Diagnosis and the Rôle of Neural Networks

Robert F Harrison
Department of Automatic Control and Systems Engineering, University of Sheffield
Sheffield, UK

R Lee Kennedy
Department of Medicine, University of Edinburgh
Edinburgh, UK

Stephen J Marshall
EDS Scicon
Camberley, UK

Research Report No 516

1. INTRODUCTION

Diagnostic problem solving, whether it be fault-diagnosis in an engineering system or diagnosis of disease in human beings, is a prime example of decision making in the face of uncertainty. Frequently, many different outcomes may correspond to an identical set of measured data or symptoms. The converse may also be true, that any given diagnosis may correspond to a number of distinct sets of diagnostic data. In addition, the data themselves may be imprecise adding to the overall uncertainty in the reasoning process, making it probabilistic in nature. These factors can often be the cause of poor diagnostic accuracy and in part responsible for the difficulty in developing useful and usable diagnostic support systems. Furthermore, it would be unusual for diagnostic errors to be viewed as equally acceptable. For example, a large number of false alarms may be tolerable in the diagnosis of heart attack when the decision to be made is simply admit to hospital or not. The level of acceptability changes though, when the decision to be made is whether or not to administer potentially life-threatening drugs. Evidently the risk associated with an incorrect diagnosis is crucial to making a decision about treatment.

Bayesian decision theory provides the formalism needed to address the problem of making risk-sensitive decisions under uncertainty and leads to the minimum risk solution to this problem. Of course, a "domain expert" must assess beforehand the relative risks or costs of making incorrect decisions . Although the optimal Bayesian classifier provides the "best" answer to any statistical decision making problem, it suffers from one serious drawback. That is, in order to develop a system, certain probability distribution functions (or probability density functions) must be obtained. This is typically done by:

i) assuming a distribution function, which may lead to inaccuracies if this does not match the true distribution;

ii) estimating the distribution function directly from data, which typically requires a large amount of data and is computationally intensive, frequently to the point of being unworkable.

This second approach is often modified by certain naive assumptions about the independence of data items, which again may lead to poor performance. It is at this stage that artificial neural networks can be used to overcome these well-known difficulties and help to implement minimum-risk Bayesian decision theory.

Artificial neural networks are computational models of the microstructure of the brain. What distinguishes them, regardless of their accuracy as brain models, is that they rely on a large number of very simple processing units, interconnected in a complex way and operating in parallel on only local information. The strengths of the connections between processing units

SHEFFIELD UNIV. LIBRARY ST. GEORGE'S

encodes the system's long-term memory of its environment and these can be adjusted according to some "learning" rule. This leads to a distributed representation of the data and to a system which can learn by example from its environment without being explicitly programmed. It will be seen that for a wide-class of neural networks, namely the "feedforward" networks, adaptation of the connection strengths leads directly to an estimator of precisely those probability distributions required for the implementation of a minimum-risk, Bayesian classifier. Furthermore, these estimates can be made as accurate as desired yet still form a parsimonious representation of the problem with little or no prior knowledge of its probabilistic structure. By their non-linear nature, the neural networks considered make more "use" of the data set, *ie* they exploit higher-order correlations in the data, and so, experience shows, can be made to perform well with far fewer data samples than classical estimation methods. Two specific architectures are discussed and it will be seen how the minimum-risk classifier may be implemented as straightforward post-processing of the outputs of a feedforward neural network.

An example, taken from medical diagnosis—the early diagnosis of heart attack—will then be discussed. This is currently the subject of research and a neural network-based decision aid developed by the authors is presently at the clinical trial stage. This case study is used to demonstrate a methodology for developing a neural network-based decision support system.

2. BAYESIAN DECISION THEORY: A BRIEF EXPOSITION

The question of assigning differential diagnoses in some problem domain when there is significant uncertainty both about symptoms and about true outcome can be effectively dealt with within the formalism of Bayesian decision theory. It can be expressed most generally as deciding to which category, or class, a particular set of data belongs. The uncertainty in the problem is most naturally expressed in terms of probabilities and an obvious objective is to design a classifier which minimises the probability of assigning an incorrect diagnosis. However, in many situations minimizing the probability of making an error is not always sufficient since, under certain circumstances, failure to diagnose a serious defect or disease is clearly less acceptable than registering a false alarm. Evidently in safety critical systems and acute medicine the relative risk involved in making a false negative judgement is very high, whereas in less critical situations, for instance where lives are not at risk, one must weigh the cost of system downtime or needlessly used resources against the risk of failure. Therefore, in developing a diagnostic aid it would be advantageous to be able to bias the outcomes towards those which are in some sense the least risky. Bayesian decision theory provides a framework which allows the risk or cost associated with particular decisions to be minimised, subject only to the availability of certain statistical information about the problem.

In this section we outline the theory of the Bayesian method and discuss the concept of relative risk. We then discuss the practicalities of implementing an optimal Bayesian classifier directly and, from there, motivate the use of artificial neural networks to perform the required estimation.

2.1 The Bayesian Formulation

Let us assume that we have M classes, $c_i \; 1 \le i \le M$, corresponding to M regions, $X_i \; 1 \le i \le M$, in the N-dimensional data-space, X. Furthermore, we assume that the regions X_i are disjoint and that they cover the entire space of all possible data, $X = \bigcup_{i=1}^{i=M} X_i$. This says that our data (written as an N-vector of elements) is drawn from an environment, X, which is partitioned into exactly M non-overlapping regions X_i, each of which is assigned a class, c_i. This ensures that every possible data vector must yield a unique diagnosis and has a diagnosis associated with it. The problem is now one of assigning a particular data vector to its proper class. We define a decision rule $d(x) = d_i$ if c_i is true (ie $x \in X_i$) assuming the obvious pairing of decision and class. Should the decision d_i be made when c_j is true and $i \ne j$ an incorrect decision is made and the data are misclassified. We therefore wish to weight decisions, d_i, and classes c_j, in such a way as to reflect the cost or risk of making a particular decision. We do this by assigning to each pair (d_i, c_j) a unique risk, $\rho_{ij} \ge 0$, *ie* ρ_{ij} is the risk involved in deciding d_i when the data belong to class c_j . Frequently it is assumed that there is no risk involved in making a correct decision although it is acceptable to assign such a risk. Indeed if full economic costs are to be accounted for then it makes sense to weight all possible pairings.

Denoting the joint probability of the occurrence of two events *a* and *b* by *P(a,b)* we define the average risk, $\Re$, as

$$\Re \overset{\Delta}{=} \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} \rho_{ij} P(d_i, c_j) \tag{1}$$

but from Bayes theorem $P(a,b) = P(a|b)P(b)$ where $P(a|b)$ denotes the probability of event *a* conditioned on event *b*. Equation (1) therefore becomes

$$\Re = \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} P(c_j) \rho_{ij} P(d_i|c_j) \tag{2}$$

If the data $x \in X_i$ then d_i is decided so that $P(d_i|c_j) = P(x \in X_i|c_j)$ or

$$P(d_i|c_j) = \int_{X_i} p(x|c_j)dx \tag{3}$$

Here, *p*(.) denotes a probability *density* function. Substituting (3) into (2) we get

$$\begin{aligned} \Re &= \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} P(c_j) \rho_{ij} \int_{X_i} p(x|c_j)dx \\ &= \sum_{i=1}^{i=M} \int_{X_i} \sum_{j=1}^{j=M} P(c_j) \rho_{ij} p(x|c_j)dx \end{aligned} \tag{4}$$

assuming the interchangeability of summation and integration. Minimisation of the average risk, $\Re$, is equivalent to choosing the regions, X_i, in equation (4) so that $x \in X_i$ if

$$\sum_{j=1}^{j=M} \rho_{ij} P(c_j) p(x|c_j) < \sum_{j=1}^{j=M} \rho_{kj} P(c_j) p(x|c_j) \;\; \forall k \neq i \tag{5}$$

The decision rule therefore becomes $d(x) = d_i$ if inequality (5) is satisfied *ie* we assign the current data vector, x, to class *i*..

We note that the special choice of weights

$$\rho_{ij} = \begin{Bmatrix} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{Bmatrix} \tag{6}$$

leads to the classifier which minimises the probability of making an error, *ie* all incorrect decisions are equally important and there is no risk involved in making a correct decision. This equates with the Maximum A Posteriori (MAP) decision criterion which states that, given the data, x, chose the class with the maximum posterior probability. This is easily seen by substitution of condition (6) into inequality (5).

2.2 Estimating the Required Probabilities

From condition (5) we see that in addition to specifying the weighting, ρ_{ij}, we also require the prior probabilities of each class and the probability density function of the data conditioned on class membership for each class. Estimation of the former is straightforward and assuming large enough data samples can be obtained from the relative frequency of occurrence of data with known classification. The structure of the conditional probability density function, $p(x|c_j)$ $1 \leq j \leq M$, is more problematic, requiring very large quantities of data which require intensive processing. We therefore appeal once more to Bayes theorem and note that,

since $$p(x|c_j) = \frac{P(c_j|x)p(x)}{P(c_j)} \tag{7}$$

inequality (5) may be restated as

$$\sum_{j=1}^{j=M} \rho_{ij} P(c_j|x) < \sum_{j=1}^{j=M} \rho_{kj} P(c_j|x) \;\; \forall k \neq i \tag{8}$$

again d_i is decided if (8) is satisfied. To implement (8) we no longer have to estimate probability *density* functions but are still faced with the task of having to estimate the probabilities of class membership conditioned on the data—the so called posterior probabilities. For high dimensional problems this is again a daunting task and has led to a commonly used assumption—that the components of the data vector are independent of one another. This assumption reduces the computational complexity of the problem and results in a reduction from the estimation of 2^{M+N} probabilities to the estimation of only $M \times N$ with the concomitant reduction in the required amount of data[1]. It is clear that any particular problem will be unlikely to satisfy this naive assumption and frequently will violate it strongly, leading to poor classification performance.

An alternative approach to direct estimation of the posterior probabilities is to introduce a parametric model of the underlying distribution which can then be adjusted to give the best (in some sense) fit. This is an approach frequently adopted in textbooks where it is often assumed that the underlying distribution is jointly Gaussian. Such an approach is useful in demonstrating the limits of performance of Bayesian classifiers and of course applies when the processes to be classified are nearly Gaussian. In many situations this will not be the case and a more generally applicable approach is needed. It is then that a "sufficiently rich" form of distribution must be chosen so that data regression methods will yield a good fit. The main difficulty here lies in the prior choice of such a distribution. Commonly a linear combination of orthogonal functions is chosen[2], but even here the choice of functions and the number of terms must be made in advance and is not guaranteed to represent the data well.

In the following section we examine the possibility of using artificial neural networks to provide a parametric model of sufficient generality that it applies to any given classification problem. The advantage here is that, although some a priori design choices must be made, they are simple to understand and therefore open the field of developing a diagnostic support system to "domain experts" rather than experts in statistical decision theory.

3. CLASSIFICATION BY NEURAL NETWORKS

In this section we are concerned with a particular class of neural network—the so called feedforward networks—which can be shown to possess attractive properties when considered within the framework of Bayesian decision theory. We first ask the question "what is a neural network and how does it differ from conventional computing?".

A neural network comprises a set of very many primitive processing elements (neurons) which each process signals in a "simple" way[a], all operating in parallel on purely local information. By contrast computers based upon the von Neumann architecture consist of one (or a small number) of highly complex processors which process data sequentially according to a pre-defined algorithm. The processing power of the neural network derives from the complex patterns of connectivity between neurons. Furthermore in a neural network, memory is instilled into the strength of these interconnections and is thus distributed throughout the network: contrast this with the single address-based memory of a von

[a] In the animal brain neuronal activity is in fact governed by nonlinear partial differential equations and this has the potential for highly complex dynamic behaviour. Here by "simple" we mean in the sense of overall input to output behaviour.

Neumann machine. At the time of writing neural networks must be simulated on conventional (serial) processors since hardware implementations are yet to become commercially available.

As far as using neural networks to alleviate the implementational problems of Bayesian decision theory goes, the link between neural networks and brain modelling is not significant. What is of interest is:
1) certain results which have been proved by the neural network community about the capacity for a feedforward neural network to represent mappings or functions and
2) that when a feedforward network adapts to (learns from) its environment it is in fact estimating the required probability distributions for Bayesian reference.

We therefore make use of a feedforward network which implements a function from the data-space (N-dimensional) into the category space (M - dimensional), which is made up of artificial neurons (units) arranged in layers according to Figure 1.

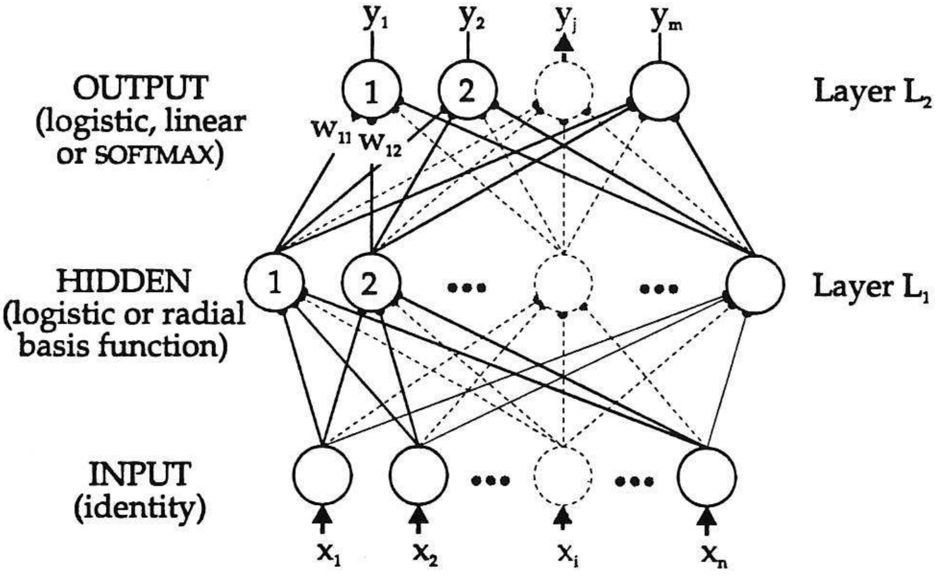

Figure 1: A two layer feedforward network;
further hidden layers may be added as required.

Here, the output of each unit is some function of its weighted (or net) input, with

$$net_i \overset{\Delta}{=} \sum_{j \in L_\ell} w_{ij} y_i \;\; ; \;\; i \in L_\ell \tag{9}$$

The weights w_{ij} determine the strength of connections between the *i*th unit in layer ℓ and the *j*th unit in layer $\ell - 1$. y_j denotes the output of the *j*th input and this is given by

$$y_i = f_i(net_i) \tag{10}$$

where f_i is some suitably smooth function usually (but not always) mapping the unrestricted input, net_i, to a finite interval. We might then write the mapping more concisely in the form

$$Y = F(W, x) \tag{11}$$

where $Y = \{y_i : i \in L_L\}$, that is Y is the vector of outputs of the network, $W = \{w_{ij}\}$, the set of all weights in the network and $x = \{x_i : 1 < i < N\}$, the data vector. In the feedforward neural network paradigm the output of the network, Y, is computed for every input in the data sample and compared with the known classification. Since the "best" value for *W* is not known a priori, errors will be generated at the output and *W* is adjusted in some way to reduce these errors. This is the adaptation or learning phase. When the errors are acceptably low (according to some measure) the network is said to be trained and learning is suppressed. The trained network can then be put into operation.

3.1 What Can a Feedforward Neural Network Compute?

A question which naturally arises is this "Is a network of the form shown in Figure 1 able to represent an arbitrary mapping from input to output?". There are two results pertaining to this depending on the precise form of the functions f_i and the structure of the network. The first is due to Cybenko[3] and states that, for the choice $f_i(u) = \dfrac{1}{1+e^{-u}}\ \forall i$, that is the logistic function, any sufficiently smooth mapping can be arbitrarily accurately approximated by a two layer (*ie* one output layer, one intermediate or hidden layer) network of the form shown in Figure 1. In this case it may be necessary to take a very large number of units in the intermediate stage and the mapping may be more parsimoniously represented by a network with more than two layers.

The second result is due to Park and Sandberg[4] and states that for a two layer network with linear output units and "radial basis functions" (radially symmetric in their arguments *eg* the Gaussian function) in the hidden layer, any sufficiently smooth mapping can be arbitrarily closely approximated.

Both of these results are most encouraging in that if we know how to choose the right structure and weightings we can implement any reasonable function by $F(W,x)$. We shall see in the following section that there is more than one way of choosing a measure of the output error for the neural network and that minimising these leads to a final structure, $F(W^*,x)$, whose *j*th element, $F_j(W^*,x)$, is an estimate of $P(c_j|x)$ required to implement the decision rule (8).

3.2 What Can a Feedforward Network Learn?

Now that we are considering the neural network as a parametric approximator to a mapping, we need to decide upon a measure of how well $F(W,x)$ approximates the desired function. It is usual in a neural network experiment to attempt to minimise some measure of the output error with respect to the weights. For instance Rumelhart *et al*[5] proposed that the mean-square output error, averaged over all samples in the training set would be a suitable measure of error and that the minimisation should be performed by gradient descent. This approach was not originated by them, but their paper certainly served to popularise this form of neuro-computing and established the so-called back propagation learning rule and its variants.

Let us assume that we have a set of representative data samples $\{x(p)\}$ where p indexes the sample over the set. Let $Y(p)$ denote the network output vector when the pth pattern is at the input and $d(p)$, the desired output for that pattern. For 1-from-M classification of an input belonging to the ith category, $d(p)$ is equal to 1 in the ith element and 0 elsewhere. The mean-square-error measure, E_{MS}, is then given by

$$
\begin{aligned}
E_{MS} &= \frac{1}{2}\sum_{p}\sum_{k=1}^{k=M}\left(d_k(p)-y_k(p)\right)^2 \\
&= \frac{1}{2}\sum_{p}\sum_{k=1}^{k=M}\left(d_k(p)-F_k(W,x(p))\right)^2
\end{aligned} \tag{12}
$$

It has been shown by numerous authors[6–9] that the optimal set of weights $W = W^*$ which minimises E_{MS} is precisely the one which yields the best (least-squares) approximation of $P_i(c_j|x)$ by $F_j(W^*,x)$ $1 \le j \le M$.

It is also possible to use other information theoretic measures of output error such as the cross-entropy function (equivalently maximum mutual information and Kullback-Leibler distance)

$$
E_{CE} = -\sum_{p}\sum_{k=1}^{k=M} d_k(p)\log F_K(W,x(p)) + (1-d_k(p))\log(1-F_k(W,x(p))) \tag{13}
$$

As before it can be shown that the optimal set of weights leads to an estimate of the required class probabilities[8–10].

Adopting (13) has the effect of weighting errors most heavily when actual outputs are close to zero or to one. Contrast this with (12) which weights large errors most heavily. This distinction, in turn, leads to the idea that the least-mean-square criterion approximates the Bayes optimal decision rule (minimum probability of error) best when $p(x)$ is large, that is, for frequently occurring data vectors. For the purposes of classification though, the best fit is required close to the decision boundaries and these will, in general, lie in regions where $p(x)$ is small[7]. By contrast the information theoretic error measure (13) is seen to be sensitive to probability estimates close to zero[9] (due to the effect of the logarithm). Such a distinction may lead one to conclude that using (13) as the objective function upon which to base an adaptation rule would lead to better overall performance. Significant improvements in classification performance have yet to be observed in practice.

A further important point in the discussion of the choice of objective function is the fact that although optimisation of E_{MS} or E_{CE} leads to feedforward networks which estimate the class conditional probabilities, neither guarantees that these estimates obey the axioms of probability. Evidently, for an output stage comprising logistic units each output must lie on the interval (0,1). It is clear that in a two class problem one output unit is sufficient and if of the logistic type then there is no such problem. In the 1-from-M (M>2) case the sum over the output is not guaranteed to equal one, so that the probability of the certain event may be greater than or less than unity. Wan[6] has considered normalising the outputs, as a post processing stage, to overcome this, on the grounds that it does not affect rank order, hence

$$w_{ij}(t+1) = w_{ij}(t) - \eta(t)\frac{\partial E}{\partial w_{ij}} \tag{15}$$

for some *i,j* and *E* the chosen objective function. The argument t indexes the passage of time and may indicate either each sample presentation or each cohort. The parameter $\eta(t)$ is a possibly time varying scalar which dictates the size of each downward step.

It should be noted that there are myriad variations on the theme of gradient descent but that these are only able to offer improvements upon, not solutions to, the problem of local structure. We therefore discuss only the most basic form. Consider now the diagram in Figure 2 which is a one-dimensional representation of the optimisation problem at hand.

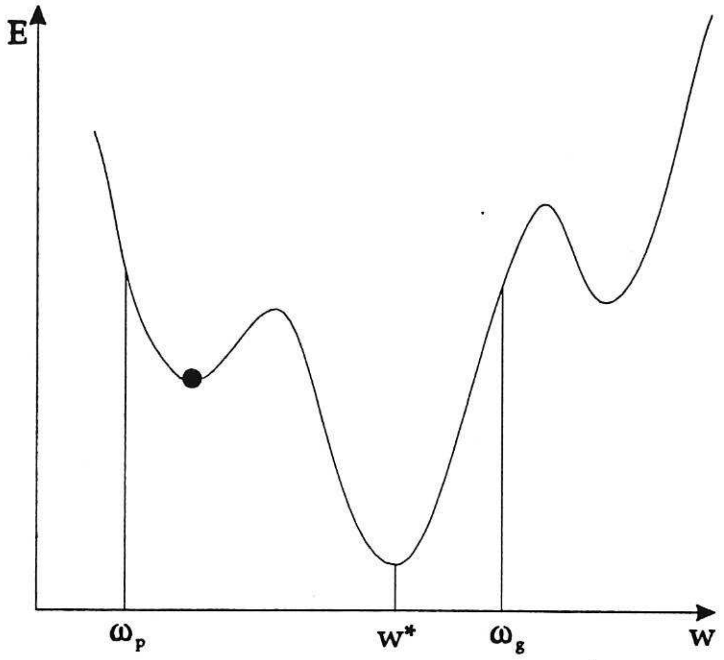

Figure 2: The objective function, *E*, as a function of connection strength, w — the scalar case

Clearly for the initial choice of $w(0) = \omega_p$ repeated downward steps (assuming a decreasing learning rate) will lead to a local minimum (marked •). For the choice $w(0) = \omega_g$ the algorithm (15) leads directly to the global minimum with corresponding weights w^* — this is the point required for the probabilistic interpretation, given earlier in this section, to hold. So, for a general feedforward structure we are faced with the problem of making a good initial choice of weights (as is so often the case in nonlinear numerical algorithms) as global information is denied to us.

The so-called radial basis function[13] feedforward structure is able completely to overcome this problem by performing classification in two stages. The radial basis function network consists of a layer of nonlinear "radial basis function" units whose outputs are combined

4.1 Classification Performance

The use of the measures E_{MS} or E_{CE} in section 3 provides a useful means of driving the learning process in feedforward neural networks. However, their use to analyse classification performance is not recommended. Certainly a small value of the chosen objective function evaluated over previously unused data will indicate that the classifier is performing well in some sense, but it is not possible to tell from this whether the system makes very many small errors (which may not affect classification performance at all) or whether it makes fewer large errors (with a consequent number of misclassifications). There are a number of performance measures which may be adopted all of which embody the same information. We shall concern ourselves here with the sensitivity, specificity and accuracy of diagnosis. These are defined as follows:

- sensitivity is defined as the ratio of the number of correct positive diagnoses to the total number of occurrences of a condition
- specificity is defined as the ratio of the number of correct negative diagnoses to the total number of non-occurrences of a condition
- accuracy is defined as the ratio of the number of correct diagnoses (both positive and negative) to the total number of cases considered.

In order to obtain a graphical picture of the trade-offs between sensitivity and specificity the Receiver Operating Classification (ROC) curve is a useful tool. This is a plot of the sensitivity *vs* 1-specificity[b] parameterized by the diagnostic threshold. Obviously, these values are calculated over an independent set of data. The typical form of the ROC curve is shown in Figure 3.

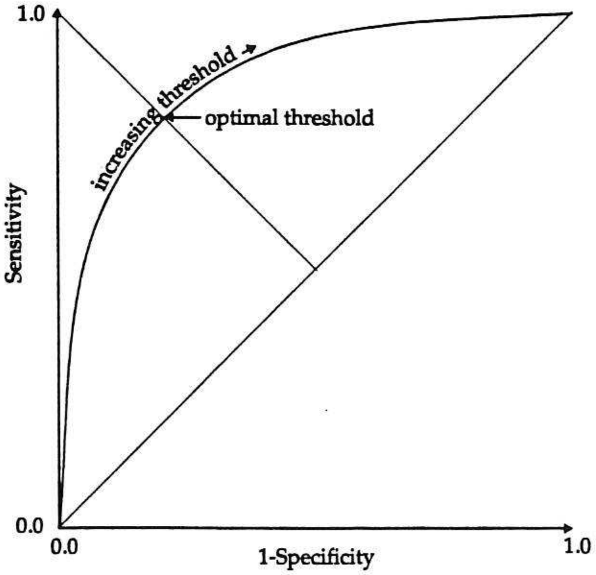

Figure 3: A typical ROC curve

[b] In decision theory sensitivity is known as the probability of detection and specificity as 1-probability of a false alarm.

Choosing a large value of threshold (top right of diagram) results in a highly sensitive classifier while a low value results in a highly specific one. The choice of threshold at the intersection of the ROC curve with the leading semi-diagonal is optimal in the sense that here accuracy = sensitivity = specificity. Multiple ROC curves can be plotted for one-from-M problems.

In medical terms, say, sensitivity therefore compares the number of people who are diagnosed as being ill with those who actually are ill and specificity the converse. It is easy, therefore, to construct a highly sensitive system since, if all patients were diagnosed as having the disease, sensitivity would be 100% but specificity would be 0%. Similarly if none of the patients was diagnosed as having the disease then specificity would be 100% and sensitivity 0%. A good predictive system is therefore one which is both highly specific and highly sensitive, that is, it would indicate those and only those patients requiring treatment. Clearly there is a trade-off between sensitivity and specificity which may be related to the relative risk of making incorrect decisions as described in section 2. Such questions are problem specific and may be considered to be design degrees of freedom.

4.2 Contribution Analysis

Neural networks are frequently criticised for their "black-box" nature, that is that their decision making is not readily inspected or understood by the user. This has led to unfavourable comparison with expert systems in particular. It would undoubtedly be advantageous to provide information on the reasoning process to an operator, in particular, on how a diagnosis has been arrived at. In single layer networks the strength of connection between data item and output unit gives a direct indication of the importance of that item in assigning a diagnosis. It also indicates that item's importance for the entire population. In the multi layer case the situation is much more complex and the value of the individual weights do not have a clear interpretation. Indeed, in general, a particular weight will play a significantly different rôle in the diagnosis depending upon the specific input data.

By calculating the sensitivity of the network outputs to the presence or absence of any data item (not to be confused with the sensitivity measure in the previous section) we are able to obtain an indication of the contribution to the diagnosis of that particular item[15,16]. We define the sensitivity of output *i* to a change in input *j*, for the p th input record, as $S_{i,j}^{p} \overset{\Delta}{=} \frac{\partial F_i}{\partial x_j}\left(x(p)\right)$. The expression $S_{i,j}^{p}$ is straightforward to develop and does not add substantially to the computational burden. This information could be used on-line to indicate to the operator the critical components in the diagnosis. It can be used off-line to obtain statistical information about the importance of particular data items in particular diagnoses which may then be compared with existing expert knowledge or, indeed, to prompt investigation into causal links.

5. A CASE STUDY IN MEDICAL DIAGNOSIS

The intention of this section is to present a case study in the early diagnosis of myocardial infarction (MI), commonly known as heart attack. Before doing this we draw attention to the use of neural networks for fault detection and diagnosis in some other fields. In the following brief discussion the neural networks are all of feedforward type and, although not explicitly treated as such, they are being used to implement Bayesian classifiers. Naidu *et al*[17], Venkatasubramanian *et al*[18], Hoskins and Himmelblau[19], Watanabe *et al*[20] and Ungar *et al*[21], have addressed the question of fault diagnosis in highly nonlinear chemical process plant with multiple symptom to multiple fault characteristics, in particular, the diagnosis of sensor failures is of interest. In a further paper Venkatasubramanian presents a case study of a fluidised catalytic cracking process and compares the effectiveness of the neural network based approach to that of a knowledge-based system[22]. In power generation, Ebron *et al*[23] propose the use of neural networks for the detection of incipient faults in an electric power distribution feeder system while Alguindigue *et al*[24] use the approach to detect changes in the state of a commercial pressurized water reactor. In other areas such as integrated circuit manufacture (Meador *et al*[25]), electronic circuit boards (Kagle *et al*[26]), or in aerospace (Barron *et al*[27], Macduff and Simpson[28], Duyar and Merrill[29], Solorzano *et al*[30]) neural networks have been used as an alternative to traditional statistical pattern recognition approaches or as a way of overcoming the inherent "brittleness" of knowledge-based approaches. The common finding in all of the above is that neural networks can offer significant advantage over existing techniques. How much advantage is still a matter for research, but the overriding message is one that is very encouraging.

5.1 The Diagnosis of Acute Myocardial Infarction

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 heart attacks 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 department, 12% of patients were found to have been erroneously discharged while 16% were found to have been inappropriately admitted to ITU[31]. In the United States approaching half of those patients admitted to ITU may ultimately be found not to have suffered a heart attack[32]. Evidently this is a diagnostic question with significant health risk and resource implications. Collinson[33] has estimated that an early transfer of patients from ITU to a medical ward may result in a financial saving of 50%. Clearly early discharge of those with non-threatening disease would result in concomitantly higher savings.

There is further pressure to improve the early diagnosis of MI, that is the advent of thrombolytic therapy which is most beneficial when administered as soon after the onset of symptoms as possible. This is because MI is caused by a blood clot which cuts off the blood, hence oxygen, supply to the coronary muscle with the resulting death of heart tissue. Thrombolytic agents are enzymes which dissolve the blood clot, unblocking the affected arteries and thereby minimising the damage to the heart. The expected benefit of thrombolytic therapy is therefore a reduction in the immediate threat to life plus an improved long term prognosis. The first of these benefits has been confirmed[34] but it is too early to make a judgement on the second. The agents must be administered within 6 hours of the infarction taking place to be of significant benefit[35]. Furthermore, they are expensive and may be dangerous if given inappropriately.

The routine diagnosis of MI relies on serial measurements of enzyme levels in the blood and the electrocardiograph. Both of these indicators rely on sufficient time having elapsed for diagnostic changes, due to the death of tissue, to have taken place. These changes may take 24–48 hours to become evident and may not be useful in the early management of the problem.

The desirability of a diagnostic aid during the early stages of MI is therefore clear, particularly for clinicians in non-cardiac specialities such as accident and emergency departments or in general practice. In what follows we demonstrate that a neural network may be used to implement a Bayesian classifier with impressive discriminating power.

5.2 Methods

5.2.1 Data collection

Acute chest pain is a suitable domain in which to develop a decision support system because of (i) its high incidence, so that data is readily available, and (ii) the possibility of obtaining a concrete final diagnosis, to supply the desired outcome to the network. Our initial study included 300 consecutive emergency referrals, with a complaint of chest pain, to a large teaching hospital in the UK[15,36]. Information was recorded from each patient on a standard pro forma, comprising 78 items of demographic, clinical and electrocardiographic data. In addition the admitting clinician was asked to estimate the likelihood (expressed as a percentage) of the patient having suffered a heart attack. Each pro forma was completed before the results from confirmatory tests were available.

Thirty-eight features were abstracted from each patient record and these were coded as a 53-dimensional bipolar vector, with zeros indicating missing data. The target or desired response was simply coded as 1 for MI and 0 otherwise. Continuous valued variables such as age and duration of pain were coded, eliminating redundant elements, using the method of Widrow *et al*[37].

5.2.2 Network architecture and training

An architecture with only one intermediate layer of units, as per Figure 1, has been used throughout. This layer is fully interconnected with the units of the input and output layers. Within layer connections and direct input-to-output couplings are not permitted. The intermediate and output layers comprise logistic units.

The final architecture, comprising 53 input units, 18 intermediate units and 1 output was arrived at after extensive experimentation[15]. Our initial configurations were guided by the algorithm[14].

In this study the mean-square objective function, E_{MS}, has been chosen to derive the optimal weight values. The network was trained on the first 90 patient records until the mean-square error was at an acceptably low value. The remaining 210 patterns were used to assess the network's diagnostic capability.

12. Yair E and Gersho A, The Boltzmann Perceptron Network: a soft classifier, *Neural Networks*, 3, 203, 1990.

13. Moody J and Darken C, Fast learning in networks of locally-tuned processing units, *Neural Computation*, 1, 281, 1989.

14. Gutierrez M, Wang J and Lipschik G, Estimating hidden units for two-layer perceptrons, in *Proc IEE 1st Int Conf on Artificial Neural Networks*, London, 1989, 120.

15. Harrison R, Marshall S and Kennedy R, A connectionist aid to the early diagnosis of myocardial infarction, in *Proc 3rd European Conf on AI in Medicine*, Maastricht, 1991, 119.

16. Marshall S, Harrison R and Kennedy R, Neural classification of chest pain symptoms: a comparative study, in *Proc 2nd IEE Int Conf on Artificial Neural Networks*, Bournemouth, 1991, 200.

17. Naidu S R, Zafiriou E, McAvoy T, Use of neural networks for sensor failure detection in a control system, *IEEE Control Systems Magazine*, 10, 49, 1990.

18. Venkatasubramanian V, Vaidyanathan R, Yamamoto Y, Process fault detection and diagnosis using neural networks: I. Steady State Processers, *Computers and Chemical Engineering*, 14, 699, 1990.

19. Hoskins J, Himmelblau D, Fault detection and diagnosis via artificial neural networks, in *Computer Applications in the Chemical Industry*, Eckermann, R (Ed.),
VCKH Verlagsgesellschaft, Weinheim, 1989, 277.

20. Watanabe K, Matsuura I, Abe M, Kubota M, Himmelblau D, Incipient fault diagnosis of chemical processes via artificial neural networks, *AIChE Journal*, 35, 1803, 1989.

21. Ungar L, Powell B and Kamens S, Adaptive networks for fault diagnosis and process control, *Computers and Chemical Engineering*, 14, 561, 1990.

22. Venkatasubramanian V, King C, A neural network methodology for process fault diagnosis, *AIChE Journal*, 35, 1993, 1989.

23. Ebron S, Lubkeman D and White M, A neural network approach to the detection of incipient faults on power distribution feeders, *IEEE Transactions on Power Delivery*, 5, 905, 1990.

24. Aguindigue I, Eryurek E, Upadhyaya B and Uhrig R, Using artificial neural networks to identify nuclear power plant states, *Transactions of the American Nuclear Society*, 61, 217, 1990.

25. Meador J, Wu A, Tseng C, Lin T, Fast diagnosis of integrated circuit faults using feedforward neural networks, in *Proceedings International Joint Conference on Neural Networks*, 1, IEEE, 1991, 269.

26. Kagle B, Murphy J, Koos L and Reeder J, Multi-fault diagnosis of electronic circuit board using neural networks, in *International Joint Conference on Neural Networks*, 2, IEEE, 1990, 197.

27. Barron R, Cellucci R, Jordan P, Beam N, Hess P and Barron A, Applications of polynomial neural networks to FDIE and reconfigurable flight control, in *Proceedings of the IEEE National Aerospace and Electronics Conference*, 2, IEEE, 1990, 507.

28. Macduff R and Simpson P, An investigation of neural networks for F-16 fault diagnosis, II: System Performance, in *Proceedings of the SPIE—Applications of Artificial Neural Networks*, SPIE, 1990, 42.

29. Duyar A and Merrill W, Fault diagnosis for the space shuttle main engine, *Journal of Guidance, Control and Dynamics*, 15, 384, 1992.

30. Solorzano M, Ishii D, Nickolaisen N and Huang W, Detection and classification of faults from helicopter vibration data using recently developed signal processing and neural network techniques, in *Conference Record of the 25th Asilomar Conference on Signals, Systems and Computers*, 2, IEEE, 1991, 1138.

31. Emerson P, *et al*, An audit of the management of patients attending an accident and emergency department with chest pain, *Quart J Med* 70, 213, 1989.

32. Pozen M, *et al*, A predictive instrument to improve coronary care unit admission practices in acute ischaemic heart disease: a prospective multi-centre clinical trial *New England J Med* 310, 1273, 1984.

33. Collinson P, Ramhamadany E, Rosalki S, Joffe J, Evans D, Fink R, Greenwood T and Baird I, Diagnosis of acute myocardial infarction from sequential enzyme measurements obtained within 12 hours of admission to hospital, *J Clinical Pathology* 42, 1126, 1989.

34. Simoons M, Thrombolytic therapy in acute myocardial infarction, *Ann Rev Med* 40, 181, 1989.

35. McNeill A, Flannery D, Wilson C, *et al*, Thrombolytic therapy within one hour of the onset of acute myocardial infarction *Quart J Med* 79, 487, 1991.

36. Kennedy R, Harrison R, Kirklis K, Moriarty K, Gangi M, Shaukat H and Young M, Early diagnosis of acute myocardial infarction: a novel approach to decision making using neural networks, *Clin Sci* 78, suppl Abstr 88, 1990.

37. Widrow B, Gupta N and Maitra S, Punish/reward: learning with a critic in adaptive threshold systems, *IEEE Transactions on Systems, Man and Cybernetics*, 3, 455, 1973.

38. Goldman L, Weinberg M, Weisberg M, Olshen R, Cook E, Sargent R, Lamas G, Dennis C, Wilson C, Deckelbaum L, Fineberg H and Stiratelli R, A computer-derived protocol to aid the diagnosis of emergency room patients with acute chest pain, *New England J Med* 307, 588, 1982.

39. Goldman L, Cook E, Brand D, Lee T, Rouan G, Weisberg M, Acampora D, Stasiulewicz C, Walshon J, Terranova G, Gottlieb L, Kobernick M, Goldstein-Wayne B, Copen D, Daley K, Brandt A, Jones D, Mellors J and Jakubowski R, A computer protocol to predict myocardial infarction in emergency department patients with chest pain *New England J Med* 318, 797, 1988.

40. Kennedy R, Harrison R, and Marshall S, Analysis of clinical and electrocardiographic data from patients with acute chest pain using a neurocomputer, *Quart J Med* 80, 788, 1991.

41. Baxt W, Use of artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion, *Neural Computation* 2, 480, 1990.

42. Baxt W, Use of an artificial neural network for the diagnosis of myocardial infarction, *Ann Int Med* 115, 843, 1991.

SHEFFIELD UNIV. LIBRARY ST. GEORGE'S