Neural Networks, Heart Attack and Bayesian Decisions: An Application of the Boltzmann Perceptron Network

Robert F Harrison
Department of Automatic Control and Systems Engineering
The University of Sheffield
PO Box 600
Mappin Street
Sheffield S1 4DU
+44 (0)742 825139

Stephen J Marshall[1]
Department of Automatic Control and Systems Engineering
The University of Sheffield

R Lee Kennedy
Department of Medicine
The University of Edinburgh

Short Title: Diagnosis of Heart Attack by Neural Networks

Research Report No 517

[1]Supported by the Science and Engineering Research Council, UK

Abstract

A decision aid is proposed for the diagnosis of the most commonly occurring cause of emergency admission to hospital in the developed world—acute myocardial infarction, or heart attack. The motivation for the proposal lies in the Bayesian (minimum risk) decision theory which is briefly reviewed. The fact that many feedforward artificial neural networks are known to estimate the conditional class probabilities required for Bayesian decision theory is explored, and one candidate—the Boltzmann Perceptron Network—is selected as possessing the most desirable properties. A brief account of the theory (based upon the so-called Boltzmann machine) underlying this little known network is presented.

The Boltzmann Perceptron Network is trained to diagnose the presence or absence of myocardial infarction on data gathered from a large UK teaching hospital and is found to perform as well as senior registrars with specific cardiological training (diagnostic accuracy in excess of 80%). In addition, the Boltzmann Perceptron Network is found to provide greater user confidence than the multi-layer Perceptron.

Keywords: Medical diagnosis, heart attack, Bayesian classifiers, artificial neural networks, Boltzmann Perceptron Network.

1 Introduction

Clinical diagnosis is a prime example of decision-making under uncertainty. Frequently, many unconnected diseases may correspond to an identical set of diagnostic data comprising symptoms, measured data and previous history. 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 may themselves be imprecise adding to the overall uncertainty in the reasoning process. These factors can be the cause of poor diagnostic accuracy which has led to the search for clinical decision aids, both general purpose and disease specific. They are also responsible, in part, for the difficulty in developing usable support systems.

Clinical decision support systems are not yet widely used in medicine although their potential impact is increasingly being recognized. Progress in this area has faced two major problems: first, the need to overcome the deficiencies in conventional approaches to decision support systems and second, the reluctance of the medical profession to use such systems. Two approaches to decision support are being studied intensively. These are expert systems and artificial neural networks. Both expert systems and artificial neural networks attempt to perform tasks akin to human reasoning. In expert systems knowledge is encoded explicitly as a series of rules and the system models the deductive process. Learning takes place either by interrogation of an expert (the knowledge elicitation process) or by statistical analysis of a database, and requires the expertise of a knowledge engineer. Because these systems execute a large number of commands sequentially they can be slow in operation although increasing availability of fast computers makes this less of a problem. Expert shells provide a framework within which the rules are encoded and speed up the process of establishing an expert system. The main obstacles to the widespread use of expert systems lie in their static nature and in the effort needed to establish and adapt them.

By contrast, artificial neural networks attempt to model the micro-structure and function of the brain. Knowledge is represented implicitly and is distributed throughout the structure of the artificial neural network. Learning takes place within the system which uses raw data to learn associations between data items—learning by example. This information is stored as numerical weights (analogous to synaptic junctions) in the artificial neural network. Artificial neural networks are self-adaptive, they can generalize well and exhibit graceful degradation and so can cope with uncer-

tainty in data items and missing data. Furthermore, rapid re-training of a network to account for regional variations is a possibility, without the need for the intervention of a "knowledge engineer". Artificial neural networks have, however, been criticized because their operation is not explicit and they cannot "explain" their reasoning, as can rule-based systems. Nonetheless, they do show considerable promise in modelling high level cognitive processes such as those involved in making medical decisions ([1,2]).

Advances in expert systems and artificial neural networks, both individually and in the production of hybrid systems, are beginning to produce usable decision support tools. However, most studies have been laboratory-based and have failed to demonstrate their potential for practical benefit. The aim of this paper is to demonstrate that artificial neural networks, in particular the so-called Boltzmann Perceptron Network, have a rôle to play in the establishment of useful and usable tools for decision support, specifically in the diagnosis of acute chest pain.

1.1 Decision support in the diagnosis of heart attack

Acute chest pain is a suitable domain in which to develop decision support systems because of its high incidence and the increasing pressure on doctors to make accurate, early diagnoses. Suspected heart attack is the commonest reason for emergency referral to hospital in the developed world and therefore a large amount of relevant data can be amassed relatively quickly. Each year in the UK alone there are over 250,000 documented heart attacks. Furthermore, a standard criterion, proposed by the World Health Organization, is available which gives a correct diagnosis (in the sense of its being widely agreed upon) but this is not available for 24–48 hours. This "gold-standard" diagnosis can therefore be used objectively to assess the performance of decision support tools.

In a recent audit of the management of acute chest pain in an Accident and Emergency department [3] approximately 12% of patients were discharged erroneously while 16% of patients were judged to have been inappropriately admitted to the coronary care unit. In the US, where 1.5 million patients are admitted to such units each year, approximately half will finally be found not to have acute ischaemic heart disease [4].

Myocardial infarction (MI), or heart attack, is caused by a blood clot blocking the coronary arteries

thus depriving heart muscle of its blood supply. The routine diagnosis of MI is made using serial changes in serum cardiac enzymes (which are released from damaged heart muscle) and serial electrocardiographs. However, using these tests it may be up to 48 hours before a definite diagnosis is made while modern therapy for acute MI demands rapid, accurate diagnosis: thrombolytic agents, which dissolve the blood clots, reduce morbidity and decrease mortality by 30–50% in patients safely reaching hospital (for review see Simoons [5]). For maximum benefit, thrombolytic agents should be administered as early as possible and certainly within six hours of the onset of symptoms [6]. In addition to this, the agents are expensive and they may be dangerous if administered inappropriately.

In addition to the health risk posed by failure to diagnose MI promptly, economic and other factors play a significant rôle in the management of chest pain victims. Indeed, it has been estimated [7] that simply making an early transfer of a patient from a coronary care unit to a general medical ward would result in a financial saving of 50%. The saving upon early discharge of patients with non-ischaemic chest pain or angina would be concomitantly higher.

The desirability of a diagnostic aid during the early stages of MI is therefore clear, particularly for those in non-cardiac specialties such as Accident and Emergency departments, or in general practice.

1.1.1 Biochemical modelling

Approaches which have been used to improve the accurate early diagnosis of MI include imaging techniques, biochemical measures and decision support. Of these, the use of biochemical tests has found the widest usage. Rapid sequential enzyme measurement gives information on the rate of change of cardiac enzymes and can be used to predict whether the diagnostic threshold for each enzyme will be exceeded. Early work [7] indicates that high accuracy is achievable, although requiring measurements up to 12 hours post-onset of pain. Diagnoses made in this way still require some hours during which time the patient may not have received thrombolytic treatment.

confidence that may be assigned to the classification unless the network output happens to represent the posterior distribution well in the first place.

To overcome this problem so that a decision can be made which truly minimizes the Bayes risk function we propose the use of the Boltzmann Perceptron Network devised by Yair and Gersho [25]. This network possesses a Perceptron-like feed-forward structure, followed by a competitive layer which "shares out" the unit probability between the outputs. The Boltzmann Perceptron Network is inspired by and derived from the Boltzmann Machine [26] and implements an approximate, deterministic version of it. The outputs of the Boltzmann Perceptron Network may therefore be considered as estimates of the posterior probabilities of the possible diagnoses, given a specific set of symptoms and clinical data—precisely those quantities required to formulate the Bayesian risk function. The Boltzmann Perceptron Network therefore overcomes one of the major problems involved in Bayesian decision theory—the need to assume that the components of the data vector are independent in order to make tractable the estimation of the probability of class membership conditioned on the data. Using the Boltzmann Perceptron Network, we have confirmed our earlier results and are now able to classify patients into different risk groups [27].

The paper is organized as follows: In Section 2 we briefly review decision-making under risk from a Bayesian viewpoint and indicate what is required to realize a Bayesian classifier. In Section 3 we examine the shortcomings of the widely used multi-layer Perceptron as an estimator of posterior probabilities and use these to motivate the use of the Boltzmann Perceptron Network. The theoretical basis for this network is then presented. The use of the Boltzmann Perceptron Network for diagnosis of acute chest pain is described in Section 4 and our results are presented and compared with other classifiers and the performance of clinicians.

2 Bayesian decision theory

The problem of deciding, in the face of uncertainty, to which category, or class, a set of data belongs can be dealt with in many ways. Clearly, one would normally wish to minimize the probability of making an incorrect decision but in situations where the cost or risk associated with making an error is high it is desirable to weight or bias the decision towards the one which is least risky.

Bayesian decision theory allows risk to be minimized formally, subject only to the availability of certain statistical quantities. It is well known that the required quantities are difficult to obtain as their calculation is computationally intensive and requires very large data samples. For these reasons it is usual to make often quite unjustified assumptions about the probabilistic nature of the problem to facilitate computation.

Here we provide a brief exposition of Bayesian decision theory which is included to indicate our motives.

The decision problem can be stated as follows: there are M classes, c_i, corresponding to M regions, X_i $1 \leq i \leq M$, in data-space, to which the N-dimensional data vectors, $\boldsymbol{x}$, may belong. The decision rule, $d(\boldsymbol{x})$, can then be written $d(\boldsymbol{x}) = d_i$ if c_i is true (*ie* if $\boldsymbol{x} \in X_i$) assuming a natural pairing of decision and class, and that there are the same number of classes as decisions to be made. It is further assumed that the regions, X_i, are disjoint and that they cover the space of all possible data, $X = \bigcup_{i=1}^{i=M} X_i$. This ensures that each point in X yields a unique decision and has a decision associated with it.

Clearly, should d_i be decided when c_j is true and $i \neq j$ an erroneous decision will have been made. Associated with each decision d_i and c_j is a unique *cost* or *risk*, $R_{ij} \geq 0$:

$$R_{ij} = \text{the cost of deciding } \mathrm{d_i} \text{ when the data belong to class } \mathrm{c_j} \tag{1}$$

In many applications it is assumed that there is no risk involved in making a correct decision although it is perfectly acceptable to assign such a risk.

Each category, c_i has an *a priori* probability, $P(c_i)$, associated with it. Clearly, since the data vector must belong to one of the classes

$$\sum_{i=1}^{i=M} P(c_i) = 1 \tag{2}$$

The average risk, $\Re$, for this problem is defined by

$$\Re \triangleq \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} R_{ij} P(d_i, c_j) \tag{3}$$

where $P(a, b)$ denotes the joint probability of events a and b. Re-writing $P(d_i, c_j)$ as $P(d_i \mid c_j)P(c_j)$ yields

$$\Re = \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} P(c_j) R_{ij} P(d_i \mid c_j) \tag{4}$$

where $P(a \mid b)$ denotes the conditional probability of events a and b. If the data $x \in X_i$ then we decide d_i, hence $P(d_i \mid c_j) = P(x \in X_i \mid c_j)$ or

$$P(d_i \mid c_j) = \int_{X_i} p(x \mid c_j)dx \tag{5}$$

where $p(.)$ indicates a probability density function. This makes the average risk

$$\begin{aligned} \Re &= \sum_{i=1}^{i=M} \sum_{j=1}^{j=M} P(c_j)R_{ij} \int_{X_i} p(x \mid c_j)dx \\ &= \sum_{i=1}^{i=M} \int_{X_i} \sum_{j=1}^{j=M} P(c_j)R_{ij}p(x \mid c_j)dx \end{aligned} \tag{6}$$

assuming the interchangeability of summation and integration. Minimizing the average risk, $\Re$, is equivalent to choosing the regions, X_i, in equation (6) so that $x \in X_i$ if

$$\sum_{j=1}^{j=M} R_{ij}P(c_j)p(x \mid c_j) < \sum_{j=1}^{j=M} R_{kj}P(c_j)p(x \mid c_j) \quad \forall k \neq i \tag{7}$$

Inequality (7) forms the decision rule in that $d(x) = d_i$ if it is satisfied. It should be noted that, under the assumptions that there is no risk associated with a correct decision and that the risks associated with making erroneous decisions are equal, the decision rule reduces to one which minimizes the probability of making an error.

One final application of Bayes' theorem puts (7) into a more convenient form for our purpose. Noting that

$$p(x \mid c_j) = \frac{P(c_j \mid x)p(x)}{P(c_j)} \tag{8}$$

and re-arranging, (7) is re-stated in the form

$$\sum_{j=1}^{j=M} R_{ij}P(c_j \mid x) < \sum_{j=1}^{j=M} R_{kj}P(c_j \mid x) \quad \forall k \neq i \tag{9}$$

and d_i is decided if (9) is satisfied. The advantage of this form from a practical point of view is that it does not require the estimation of the conditional probability *density* functions—a notoriously difficult task especially in high dimensional data, requiring very large data samples. As it is, the probabilities of class membership conditioned on the data must be estimated to obtain a viable decision rule. This is very frequently an impossible task for practical reasons and has led to an assumption which is commonly used—that the components of the data vector are independent. This reduces the computational complexity of the estimation problem to one of estimating each distribution conditioned on only one element of x. The computational benefit is a reduction from

the estimation of 2^{M+N} probabilities to $M \times N$ [28, *pp* 266–268], leading to a concomitant reduction in the necessary quantity of data. Equally clear, however, is the fact that any given problem will not fulfil this assumption and frequently will violate it strongly, leading to a classifier which performs poorly. In the following section we examine the possibility of using neural networks for estimating the posterior probabilities directly, without making such assumptions.

3 Estimation of posterior probabilities by neural networks

3.1 The multi-layer Perceptron

We briefly summarize the conclusions of [24,29] which describe how the widely used multi-layer Perceptron [21] relates to the estimation of posterior probabilities. Familiarity with the multi-layer Perceptron is assumed. Both articles point out that it is the minimization of a least-mean-square error criterion which leads to the Bayesian interpretation and that any structure capable of describing the mapping from the input (data) space to the output (probability) space will be valid. As is well known, the multi-layer Perceptron with adequate numbers of hidden units and/or layers can form arbitrary input/output maps [30] and is thus a good candidate. We shall not address the problem of how to decide on the optimal network structure.

Ruck and colleagues [29] demonstrate that the multi-layer Perceptron approximates the Bayes optimal decision rule (minimum probability of error) for both binary and multiple decisions. They point out that the multi-layer Perceptron fits the optimal discriminant function best when $p(x)$ is large, that is, for frequently occurring data vectors. However, for the purposes of classification, the best fit is required close to the decision boundaries and these will, in general, lie away from the maximum of $p(x)$. The proof in [29] also indicates that the outputs of a trained multi-layer Perceptron may be interpreted as estimates of the $P(c_j \mid x)$'s. This result is also derived in [24] where it is shown that, although the outputs are estimates of the posterior probabilities, they do not in general possess the properties of probability distributions; in particular, they do not sum to one. Output normalization is proposed to overcome this problem on the grounds that it does not affect rank order and hence classification. However, it does affect the level of confidence associated with a classification. The author goes on formally to quantify this.

By adopting the Boltzmann Perceptron Network we propose to overcome this problem since the network's estimates of the $P(c_j \mid x)$'s satisfy the axioms of probability theory directly.

3.2 The Boltzmann Perceptron Network

The Boltzmann Perceptron Network is a deterministic feedforward network. It is functionally equivalent to the Boltzmann Machine in that it computes class-conditional probabilities, yet it has a deterministic feedforward structure similar to the multi-layer Perceptron. The network is composed of two successive stages. In the first stage (similar to a two-layered Perceptron) decision functions are evaluated for each output class. These are fed to the second stage which performs a *soft competition* between the classes, resulting in a graded "scoring" for the different class outputs. Instead of a sole winner, the classes share the total amount of unit probability, thus guaranteeing that the outputs sum to one.

The Boltzmann Perceptron Network is essentially a deterministic model of the Boltzmann Machine at "thermal" equilibrium. The steady-state properties of the Boltzmann Machine are central to its derivation but the Monte-Carlo approach (simulated annealing) to achieving this equilibrium is not required (although the idea of an artificial "temperature" is still used). Evaluating the deterministic equations which describe the Boltzmann Machine at equilibrium is, in general, impractical since it involves complexity of order 2^{J+M} where J and M are the numbers of "hidden" and output units, respectively. However, by imposing a number of constraints on the deterministic equations (rather than on the underlying Boltzmann Machine itself), the Boltzmann Perceptron Network realizes a computational complexity of order JM, which is computationally feasible. Although these constraints cause the Boltzmann Perceptron Network to deviate from the underlying Boltzmann Machine, preliminary performance results indicate that it performs well as a (soft) pattern classifier. In common with the Boltzmann Machine the Boltzmann Perceptron Network has a learning algorithm based upon relative entropy or Kullback-Leibler number [31] rather than the least-mean-square error criterion of the multi-layer Perceptron. However, since no simulated annealing is required and because of the relatively low complexity of the equations, learning in the Boltzmann Perceptron Network is far less time consuming. We outline the key steps in the derivation of the Boltzmann Perceptron Network since it is a paradigm which has so far received little exposure in the litera-

ture. The development here follows that of Yair and Gersho [25]. First, the energy function of the Boltzmann Machine is introduced and the modifications which lead to the massive reduction in complexity are briefly described. A performance measure which is used to derive the learning algorithm is then presented.

The underlying Boltzmann Machine is symmetrically connected with no self-feedback [26] and the free running elements are allowed to change their states randomly and asynchronously. The state vector of the Boltzmann Machine is defined as $u \triangleq (x^t, v^t, y^t)$, where x is an $N \times 1$-vector of real values—a slight departure from the original conception of the Boltzmann Machine, whose values are clamped to those of the input vector (hence the equivalent notation), v is a $J \times 1$-vector containing the activations of the hidden units and y is an $M \times 1$ vector of outputs. The notation $(.)^t$ indicates the transpose operation. The "energy" function of the Boltzmann Machine may be defined thus:

$$E_u \triangleq -\frac{1}{2} u^t \Gamma u - u^t \delta \tag{10}$$

where Γ is the symmetric, zero-diagonal connection matrix and δ is a vector of biases. In view of the three distinct components of the state vector, the matrix, Γ, and the vector, δ, partition naturally, thus:

$$\Gamma = \begin{bmatrix} D_1 & R^t & W^t \\ R & D_2 & Q \\ W & Q^t & D_3 \end{bmatrix} \qquad \delta^t = \begin{pmatrix} f^t, c^t, s^t \end{pmatrix} \tag{11}$$

The energy function may then be written

$$E_u = -\left(v^t h^v_{y|x} + T_{y|x} + \eta_x \right) \tag{12}$$

where

$$h^v_{y|x} \triangleq Rx + Qy + 0.5 D_2 v + c \tag{13}$$

$$T_{y|x} \triangleq y^t (Wx + 0.5 D_3 y + s) \tag{14}$$

$$\eta_x \triangleq x^t (0.5 D_1 x + f) \tag{15}$$

$$\tag{16}$$

Since the states of the input layer are fixed (for each classification) they can be removed from the state vector which then becomes a (binary) vector of length $J + M$. This state vector and its associated energy are denoted by $v, y \mid x$ and $E_{v,y|x}$, respectively. Moreover, since η_x is not now a

function of state, but serves only as a fixed bias, it can be neglected and D_1 and f can be set to zero without loss of generality [25].

For pattern classification to be performed in a deterministic sense (rather than by accumulating statistics) it is necessary to evaluate the probability, $P(y \mid x)$ of observing a particular pattern, y, on the Boltzmann Machine given an input, x. Since the Boltzmann Machine is assumed to be in thermal equilibrium, this probability can be obtained from the Boltzmann distribution from

$$P(v, y \mid x) = \frac{1}{Z_x} e^{-\beta E_{v,y|x}} \tag{17}$$

where β is the inverse of the (artificial) temperature and Z_x is the normalizing *partition function* given by

$$Z_x = \sum_{v,y} e^{-\beta E_{v,y|x}} \tag{18}$$

Summing (17) over all permutations of v then gives the required probability

$$P(y \mid x) = \frac{1}{Z_x} \sum_{v} e^{-\beta E_{v,y|x}} \tag{19}$$

However, evaluation of the partition function is exponentially complex requiring on the order of 2^{J+M} exponentiations of the energy function.

Assuming that the set of classes is exhaustive and mutually exclusive it is possible to classify any input pattern into a single class using an output vector of the form $y = e_i,\ i \in [1, M]$, where e_i is the i^{th} standard basis vector and corresponds to class i. Only M out of a possible 2^M output vectors are actually feasible and a state vector whose output is feasible is denoted $v, e_i \mid x$. Applying $D_1 = 0$ and $f = 0$ reduces the energy equations (12) and (16) to

$$E_{v,e_i|x} = -\left(v^t h^v_{e_i|x} + T_{e_i|x}\right) \tag{20}$$

where

$$h^v_{e_i|x} \triangleq Rx + q_i + 0.5 D_2 v + c \tag{21}$$

$$T_{e_i|x} \triangleq \sum_{k=1}^{k=N} w_{i,k} x_k + s_i \tag{22}$$

$$\tag{23}$$

where q_i is the i^{th} row of Q, $w_{i,k}$ is the i, k^{th} element of W and s_i is the i^{th} element of S. Due to the zero-diagonal structure of D (and hence D_3) the product $e_i^t D_3 e_i \equiv 0 \ \ \forall i$. D_3 can therefore

take the value 0 with no loss of generality [25]. Even though the Boltzmann Machine itself is fully interconnected, the fact that D_3 does not appear in the energy equations means that there is no interaction between the processing elements of the output layer of the Boltzmann Machine when $y = e_i, \quad i \in [1, M]$ *ie* when a feasible output is generated.

By restricting the output to be feasible, the partition function may be replaced by

$$Z_x^f = \sum_{y \in e_i,\ i \in [1,M]} \sum_v e^{-\beta E_{v,y|x}} \tag{24}$$

where the superscript f indicates that a feasible output has occurred.

For the final stage of simplification (reduction in complexity) it is convenient to introduce the "score" of the i^{th} class, $L_i(x) \triangleq -\beta E_{v,e_i|x} = \beta T_{e_i|x} + A_{e_i|x}$ with

$$A_{e_i|x} \triangleq ln\left(\sum_v e^{\beta v^t h^v_{e_i|x}}\right) \tag{25}$$

so that

$$P_{e_i|x} = \frac{1}{Z_x^f} e^{L_i(x)} \tag{26}$$

and

$$Z_x^f = \sum_{e_i,\ i \in [1,M]} e^{L_i(x)} \tag{27}$$

The exponential nature of equations (26) and (27) enforces a competition between the class scores so that a relatively high score will quickly dominate and its probability will be close to unity at the expense of the other classes.

Clearly the complexity of computing $A_{e_i|x}$ is of order 2^J, however, if the matrix, D_2, is set to zero the summation in $A_{e_i|x}$ becomes separable and, defining $V_j^{e_i|x}$ by $V_j^{e_i|x} \triangleq \sum_{k=1}^{k=N} r_{j,k} x_k + c_j + q_{j,i}$,

$$A_{e_i|x} = \sum_{j=1}^{J} ln\left(1 + e^{\beta v_j^{e_i|x}}\right) \tag{28}$$

thereby reducing complexity to the order of J, which is computationally feasible [25]. With this restriction the Boltzmann Perceptron Network deviates from the underlying Boltzmann Machine in that there are now no interconnections among the units of the hidden layer. In general, this will compromise the performance of the Boltzmann Machine (and the Boltzmann Perceptron Network which models it), however, in [25] Yair and Gersho demonstrate that their computationally efficient scheme can represent posterior probabilities very accurately. Reference [25] describes in detail how

the foregoing ideas are realized as a neural network comprising two layers. The first, a layer of sigmoidal units connected in the manner of the multi-layer Perceptron. The second, a layer using the exponential function to impose a graded or "soft" competition between the classes, which shares out the available unit probability of class membership.

The parameter, β, which is the inverse of the (artificial) temperature in the Boltzmann distribution, can be adjusted to alter the sharpness between the classification boundaries—the higher the value it takes, the sharper the classification. In the limit as $\beta \rightarrow \infty$ the Boltzmann Perceptron Network becomes a maximum *a posteriori* classifier [25,32], in which the classification is made on a winner-take-all basis.

In order to enable the Boltzmann Perceptron Network to learn the desired mapping from pattern space to probability space a method of adjusting the weights and biases of the network is needed to capture the underlying, statistical constraints of the environment. The environment is specified by a set of examples, each comprising three components: a pattern vector, x; an *a priori* probability, $Q(x)$; and a set of M desired class conditional probabilities, $Q_{i|x}$ $i \in [1, M]$. For each input pattern the network produces a set of outputs, $P_{i|x}$ $i \in [1, M]$ which differs in general from the true probabilities. The learning rule is defined so as to choose the network parameters that minimize the difference (in some sense) between the $P_{i|x}$ and the $Q_{i|x}$ over the whole training set. To do this the measure of difference or distortion between the true and calculated probabilities is given by G_x, defined by

$$G_x = \sum_{i=1}^{i=M} Q_{i|x} ln \left(\frac{Q_{i|x}}{P_{i|x}} \right) \tag{29}$$

and the average distortion over the data set is therefore

$$G = \sum_x Q(x) G_x \tag{30}$$

where G is the Kullback-Leibler number [31]. The weights and biases of the network are therefore adjusted to minimize G.

For convenience the distortion, G, may be decomposed into two terms, only one of which is a function of the weights

$$G = D - H_Q \tag{31}$$

where

$$D = \sum_x Q(x) D_x \quad \text{with} \quad D_x = - \sum_{i=1}^{i=M} Q_{e_i|x} ln P_{e_i|x} \tag{32}$$

and

$$H_Q = -\sum_x Q(x) \sum_{i=1}^{M} Q_{e_i|x} ln Q_{e_i|x} \tag{33}$$

Since only D is a function of the network parameters it is this quantity which forms the cost function for the minimization. The quantity, H_Q, is an entropy measure which indicates the average uncertainty present in the data set. In [25] it is shown that H_Q is a lower bound on D and can be used to define a useful measure of network performance

$$\rho = \frac{G}{H_Q} \tag{34}$$

The search for the minimum value of ρ (equivalently G or D) with respect to the network parameters must be conducted using an appropriate numerical scheme. The partial conjugate gradient search method is chosen in [25] whereas our work has used straightforward gradient descent.

4 A diagnostic aid for acute myocardial infarction

4.1 Background

4.1.1 Patients and methods

The present study involves 500 consecutive emergency referrals with a complaint of chest pain to the Medical Department of the Northern General Hospital, Sheffield, England. Information from all patients was recorded on a standard pro-forma which consisted of 78 items of demographic, clinical and electrocardiographic data. In addition, the immediate diagnosis of the admitting clinician was recorded. Each pro-forma was completed within four hours of presentation, before the serial measurements of cardiac enzyme and electrocardiograph were available. Cardiac enzymes (creatine phosphokinase and lactate dehydrogenase) were measured each day for three days following admission and 12-lead electrocardiograph recordings were taken over the same period and during any episode of pain.

Acute MI was diagnosed if cardiac enzymes showed progressive changes in association with either a compatible clinical history, evolving electrocardiograph changes, or both. The final diagnosis was assigned independently by the senior clinician on the case, taking account of clinical, laboratory

and follow-up data. The clinician was not aware of the factors used to train the network or of its output. All patients were followed-up, and where necessary, ischaemic heart disease was excluded or diagnosed by exercise electrocardiography or coronary angiography.

The information recorded at presentation was the opinion of the medical registrar, whether or not the patient was a referral from general practice or from the casualty department. There was no statistically significant difference in age, sex, time of presentation and diagnosis between these two groups of patients.

We took 38 features from each patient record and coded them as a 53 dimensional, bipolar vector (each element taking the value -1, 0 or 1 where 0 indicates missing data) and the target vector (diagnosis) was coded as 1 for membership of a particular class, 0 otherwise. Continuous valued variables such as age and duration of pain were also coded as bipolar vectors by dividing the normal range into "bins".

The data set was divided thus: 200 patient records were used for training purposes; 100 records were used in the validation phase; and 200 records were taken to be the test set, used to evaluate network performance. The validation phase is inevitably required when developing neural network applications due to the non-linear nature of the "optimization" or training problem which may suffer from local phenomena. It is therefore necessary to train a number of network configurations and to compare their performance using previously unseen data. The network which performs best over the validation set, according to some objective criterion, is then selected and its performance is evaluated on the test set. Since the validation set is used in the establishment of the final network it is no longer valid as test data—a point frequently ignored by applications developers.

The diagnostic capability of a trained network was assessed by calculating the following performance indicators for the set of previously unseen data. They are defined as follows—*Diagnostic accuracy*: ratio of number of correct diagnoses to total number of cases; *Sensitivity*: ratio of number of correct *positive* diagnoses to the total number of patients *with* the disease; *Specificity*: ratio of number of correct *negative* diagnoses to the total number of patients *without* the disease.

Using these indicators it is easy to plot the receiver operating characteristic (ROC) curve (sensitivity *vs* 1-specificity) for a binary classifier. The intersection of the ROC curve with the semi-diagonal

indicates the optimal threshold level, T_o, (*ie* sensitivity = specificity = accuracy) that a network can operate at, for a binary decision. The operating threshold for a minimum probability of error classifier bears an interesting relationship to the *relative* risk of making an incorrect decision (assuming that the risks associated with a correct decision are zero). It is straightforward to show that any decision threshold, T, say, is related to the risk factors in the following way:

$$T = \frac{1}{1 + \frac{R_{21}}{R_{12}}} \tag{35}$$

where R_{21} is the risk associated with a *false negative* decision and R_{12} is the risk associated with a *false positive* decision. Therefore, operating at different thresholds is equivalent to assigning some relative risk to the decision.

From a user's point of view, operating at a threshold of T_o, may be counter-intuitive if T_o is not close to 50%, the "natural" threshold. Of course, depending on the risks associated with a decision, clinicians may implicitly adjust their operating characteristics but this would not be easy to quantify unless an accurate assessment of their (subjective) probability of diagnosis could be made. The neural network can, however, be adjusted to give a 50% operating point while still performing optimally by making use of equation (35). We re-write (35) thus: $r_o \triangleq R_{21}/R_{12} = (1 - T_o)/T_o$ for the optimal threshold, $T = T_o$ and weight the network outputs as follows. Let y_i $i \in [1, 2]$ denote the output corresponding to the i^{th} class, then weight y_1 by $1/(y_1 + r_o y_2)$ and y_2 by $r_o/(y_1 + r_o y_2)$. This has the effect of shifting the operating threshold to the subjectively more appealing 50% point, while ensuring that the outputs—*weighted* probabilities—still sum to one.

In this study we have not attempted to assign risks or costs to decisions although it is clear that all misclassifications cannot be regarded as of equal seriousness. Indeed, the risk associated with a false positive diagnosis, with the attendant implication that the patient will be admitted to a coronary care unit, is insignificant compared to that of a false negative and the possibility that the patient will be discharged.

4.2 Results

Using the Boltzmann Perceptron Network to classify patients into two groups according to the presence or absence of MI, the results obtained were comparable to, or slightly better than, those

obtained by the unaided clinicians (see Table 1) who were experienced registrars with specific cardiological training. The sensitivities, specificities and accuracies for the clinicians and the Boltzmann

	Clinician	BPN (50%)	BPN (26%)
True +ve	45	39	47
True -ve	118	134	124
False +ve	20	4	14
False -ve	17	23	15

Table 1: Clinicians' preliminary diagnoses *vs* Boltzmann Perceptron Network

Perceptron Network are shown in Table 2. We present results for the network operating at a 50%

	Clinician	BPN (50%)	BPN (28%)
Sensitivity (%)	73	63	76
Specificity (%)	86	97	90
Accuracy (%)	82	87	86

Table 2: Performance of clinicians' and Boltzmann Perceptron Network

decision threshold and at the optimal threshold, T_o (= 0.28). Note that T_o is computed over the *validation* data which has the effect of levelling the three performance measures. However, in operation (over the *test* data) we find that they are no longer equal (Table 2). This is due to statistical differences in these small samples. It was not possible to account for the operating characteristic of the clinicians since their diagnosis was recorded simply as the presence or absence of MI. It is clear that operating the Boltzmann Perceptron Network at T_o dramatically improves its performance.

To compare, formally, the performance of the Boltzmann Perceptron Network and the clinicians we use McNemar's test for matched-samples which is commonly used to compare different methods of making clinical decisions [33, *pp* 255–258]. By highlighting the frequency with which techniques *disagree*, a test statistic can be generated which is χ^2-distributed with one degree-of-freedom. Table 3 displays the numbers of cases for each possible decision pairing for the clinicians and for the network operating at its optimal decision threshold.

The test statistic, $\mathcal{M}$, is calculated thus: $\mathcal{M} = (|N_{pos/neg} - N_{neg/pos}| - 1)^2/(N_{pos/neg} + N_{neg/pos})$ using the information in Table 3. $N_{./.}$ denotes the number of disagreements between methods and

		Network	
		Pos	Neg
Clinicians	Pos	47	18
	Neg	14	121

Table 3: Positive and negative MI diagnoses. Clinician *vs* network

the null hypothesis is that there is no significant difference between the diagnostic performance of the network and the clinicians. Here $\mathcal{M} = 0.28$ which is not significant at the 99% level and the null hypothesis is accepted *ie* the Boltzmann Perceptron Network and the clinicians perform equally well.

We also test the performance of the multi-layer Perceptron against the Boltzmann Perceptron Network. We do this using the Wilcoxon test for matched-pairs [33, *pp* 224–227] which takes account of the magnitude of errors. When tested on the numbers of correct decisions made over the 200 patient test set the Boltzmann Perceptron Network was found to outperform the multi-layer Perceptron (with a confidence level in excess of 99.9%) even though, on the dichotomous decision, the two methods appear to perform very similarly.

The advantage of this test for deciding upon one decision rule over another lies in the levels of certainty or otherwise that may be associated with a particular decision. For instance, although we have found linear discriminant analysis to perform well in making dichotomous decisions [16], its outputs are clustered around its optimal decision threshold (an obvious consequence of linear regression) so that when it makes a correct diagnosis it is no more confident about it than when it makes an incorrect one (*ie* its errors are equally weighted). Contrast this with the behaviour of the Boltzmann Perceptron Network which makes a correct decision with small error, and makes incorrect decisions with large error *ie* its errors are weighted in favour of *correct* decisions. This means that the user can have confidence that when the Boltzmann Perceptron Network returns a probability close to one or zero it is likely to be correct and when it returns a probability in the mid-range it indicates that the diagnosis is uncertain. The Wilcoxon test, in this case, indicates that the decision of the multi-layer Perceptron is qualitatively less reliable than that made by the Boltzmann Perceptron Network.

5 Conclusions

The Boltzmann Perceptron Network has been proposed as an "inference" engine to perform the estimation of the posterior probability of class membership, conditioned on data, in a Bayesian decision support system *ie* the estimation is performed by an artificial neural network—a deterministic implementation of the Boltzmann Machine. A second candidate artificial neural network—the multi-layer Perceptron—was considered for this rôle but was rejected for its lack of a rigourous probabilistic interpretation of its outputs, which may be important in overcoming resistance to the use of computerized decision support systems. The integration of the Boltzmann Perceptron Network into a risk-sensitive decision making system is also proposed.

An example from medical diagnosis has been presented in which the proposed decision aid is shown to perform as well as experienced registrars with specific cardiological training and to perform with a higher degree of user confidence than the multi-layer Perceptron. In an emergency setting, during the critical time period, it is unlikely that an individual would be attended by such specialized clinicians, thus indicating that a neural network-based decision aid of this type may have a significant impact in the practise of "front-line" Accident and Emergency medicine.

References

[1] Bounds D, Lloyd P and Mathew B (1990) "A comparison of neural network and other pattern recognition approaches to the diagnosis of low back disorders" *Neural Networks* **3** 583–591.

[2] Mulsant B and Servan-Schreiber E (1988) "A connectionist approach to the diagnosis of dementia" *Symp on Comp Appl in Med Care* 245–250.

[3] Emerson P, *et al* (1989) "An audit of the management of patients attending an accident and emergency department with chest pain" *Quart J Med* **70** 213–220.

[4] Pozen M, *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 J Med* **310** 1273–1278.

[5] Simoons M, (1989) "Thrombolytic therapy in acute myocardial infarction" *Ann Rev Med* **40** 181–200.

[6] McNeill A, Flannery D, Wilson C, *et al* (1991) "Thrombolytic therapy within one hour of the onset of acute myocardial infarction" *Quart J Med* **79** 487–494.

[7] Collinson P, *et al* (1989) "Diagnosis of acute myocardial infarction from sequential enzyme measurements obtained within 12 hours of admission to hospital" *J Clinical Pathology* **42** 1126–1131.

[8] Emerson P, *et al* (1988) "The development of ACORN, an expert system enabling nurses to make admission decisions about patients with chest pain in an accident and emergency department" *Medical Informatics in Clinical Medicine, Nottingham, UK* 37–40.

[9] Fattu J, Blomberg D and Patrick E (1987) "CONSULT learning system applied to early diagnosis of chest pain" *Proc* 11th *Annual Symposium on Computer Applications in Medical Care* 71–77.

[10] Boissel J and Vanarie R (1977) "Système d'aide à la phase aiguë de l'infarctus du myocarde" *Proc Int Symposium on Medical Informatics, Toulouse, France* 571–583.

[11] Goldman L, *et al*, (1982) "A computer-derived protocol to aid the diagnosis of emergency room patients with acute chest pain" *New England J Med* **307** 588–596.

[12] Goldman L, *et al* (1988) "A computer protocol to predict myocardial infarction in emergency department patients with chest pain" *New England J Med* **318** 797–803.

[13] Poretsky L, Liebowitz I and Friedman S (1985) "The diagnosis of myocardial infarction by computer-derived protocol in a municipal hospital" *Angiology* **39** 165–170.

[14] Kennedy R, Harrison R, *et al* (1990) "Early diagnosis of acute myocardial infarction: a novel approach to decision making using neural networks" *Clin Sci* **78** suppl Abstr 88.

[15] Patrick E (1977) "Update on pattern recognition applied to the early diagnosis of heart attacks" *Proc Int Conf on Cybernetics, Washington DC, USA* 742–746.

[16] Harrison R, Marshall S and Kennedy R (1991) "A connectionist aid to the early diagnosis of myocardial infarction" *Proc 3rd European Conf on AI in Medicine, Maastricht, Netherlands* 119–128.

[17] Harrison R, Marshall S and Kennedy R (1991) "The early diagnosis of heart attacks: a neurocomputational approach," *Proc Int Joint Conf on Neural Networks, Seattle, USA* **1** 1–5.

[18] Hart A and Wyatt J (1989) "Connectionist models in medicine: an investigation of their potential" *Proc European Conf on AI in medicine, London* 115–124.

[19] Baxt W (1990) "Use of an artificial neural network for data analysis in clinical decision making: the diagnosis of acute coronary occlusion" *Neural Computation* **2** 480–489.

[20] Baxt W (1991) "Use of an artificial neural network for the diagnosis of myocardial infarction" *Ann Int Med* **115** 843–848.

[21] Rumelhart D, Hinton G and Williams R (1986) "Learning representations by back-propagating errors" *Nature* **323** 533–536.

[22] Meistrell M (1990) "Evaluation of neural network performance by receiver operating characteristic analysis: examples from the biotechnology domain" *Computer Methods and Programs in Biomedicine* **32** 73–80.

[23] Kennedy R, Harrison R, *et al* (1991) "Analysis of clinical and electrocardiographic data from patients with acute chest pain using a neurocomputer" *Quart J Med* **80** 788–789.

[24] Wan E (1990) "Neural network classification: a Bayesian interpretation" *IEEE Trans on Neural Networks* **1** 303–305.

[25] Yair E and Gersho A (1990) "The Boltzmann Perceptron Network: a soft classifier" *Neural Networks* **3** 203–221.

[26] Ackley D, Hinton G and Sejnowski T (1985) "A learning algorithm for Boltzmann machines" *J Cognitive Science* **9** 147–169.

[27] Marshall S, Harrison R and Kennedy R (1991) "Neural classification of chest pain symptoms: a comparative study" *Proc 2nd IEE Int Conf on Artificial Neural Networks, Bournemouth, UK*, 200–204.

[28] Lucas P and Van der Gaag L, **Principles of Expert Systems**, Addison Wesley, Amsterdam, 1991.

[29] Ruck D, *et al*, "The multi-layer Perceptron as an approximation to a Bayes optimal discriminant function," *IEEE Trans on Neural Networks*, 1, 296–298, 1990.

[30] Cybenko G, "Approximations by superpositions of a sigmoidal function," *Mathematics of Control, Signals and Systems*, 2, 303–314, 1989.

[31] Kullback S, **Information Theory and Statistics**, John Wiley, New York, 1959.

[32] Yair E and Gersho A, "Maximum *a posteriori* decision and evaluation of class probabilities by Boltzmann Perceptron classifiers," *Proc IEEE*, 78, 1620–1628, 1990.

[33] Bland M, **An Introduction to Medical Statistics**, Oxford Medical Publications, Oxford, 1987.

Robert Harrison is a Senior Lecturer and has research interests in decision and control, non-linear and stochastic systems and neural networks. He has published over forty papers in these areas.

Stephen Marshall is a graduate student whose doctoral thesis deals with the use of neural networks for clinical decision support. He has a dozen publications in the field.

Lee Kennedy is a Senior Lecturer and Consultant Physician. His research interests lie in endocrinology and biochemistry, and in clinical decision support. He has published over fifty papers in these and related areas.