On the Use of Artificial Neural Networks in a Commercial Holter Algorithm

J Pardey

Huntleigh Healthcare, Cardiology Products Division, United Kingdom

Abstract

The Medilog ADAPT analysis algorithm uses multi-layer perceptron artificial neural networks to estimate the Bayesian a posteriori *probability that each detected QRS complex is supraventricular or ventricular given measurements such as the QRS width and height, etc. It is shown that the paradigm used in Medilog ADAPT improves on the standard paradigm for neural network classifiers in two important aspects. To solve the problem of missing data the features presented to the neural network's inputs are first converted to probability estimates, so that when a feature cannot be measured it is assigned a value of 0.5. Then during neural network training the optimum neural network is the one that generates an output probability of 0.5 when its inputs are all assigned a value of 0.5. It is shown that using this neural network Medilog ADAPT is 99.9% accurate when tested against a database of eight 3-channel 24-hour ECG recordings.*

1. Introduction

The Medilog ADAPT analysis algorithm uses multi-layer perceptron (MLP) neural networks to estimate the Bayesian *a posteriori* probability that each detected QRS complex is supraventricular, N, or ventricular, V, given measurements such as the QRS width and height, etc [1].

The standard paradigm for neural network beat classification comprises a pre-processing stage of feature extraction and normalisation, a classification stage using an 'optimum' MLP architecture and weights set, and a post-processing stage to resolve any discrepancies between the timings and/or classifications of QRS complexes detected across more than one channel. Within this paradigm the optimum MLP architecture and weights set is determined during a supervised learning phase in conjunction with a criterion for when to stop learning – namely the number of passes through the training set that minimises the mean squared error (m.s.e.) between the desired and actual MLP output values on a cross-validation set. It has been shown that MLPs trained in this way estimate Bayesian *a posteriori* probabilities [2]:

$$p(C_k \mid \underline{x}) = \frac{p(\underline{x} \mid C_k)\, p(C_k)}{\sum_{i=1}^{K} p(\underline{x} \mid C_i)\, p(C_i)}$$

where $\underline{x}$ is a vector of input features $x_1, x_2, \ldots, x_n$, $p(\underline{x} \mid C_k)$ is the class-conditional probability of the feature vector $\underline{x}$ given class C_k, $p(C_k)$ is the *a priori* probability of class k, and K is the total number of classes. For the classification problem presented in this paper, K = 2 (i.e. N or V) and therefore the MLP requires only a single output unit since $p(V|\underline{x}) = 1 - p(N|\underline{x})$. Each QRS complex is then classified as N if $p(N|\underline{x}) > 0.5$ or V if $p(N|\underline{x}) < 0.5$.

Multiple training runs are performed for each MLP architecture using a different random initialisation of the weights on each run. Once training is complete the optimum MLP is selected using one of two standard selection criteria – namely the MLP that has the smallest m.s.e., or lowest classification error rate, on the cross-validation set. The classification performance of this MLP is then evaluated on an independent test set of patients whose data are not included in the training or cross-validation sets.

The paradigm used in Medilog ADAPT improves on the standard paradigm in two important aspects. First, as a final step in the pre-processing stage each feature is converted to a probability that the QRS complex is supraventricular or ventricular on the basis of this feature alone. An advantage of this is that in the standard paradigm it is not clear what value to assign to an MLP input unit when a measurement is missing or cannot be made – for example, due to noise or artefact – whereas in the new paradigm such an input unit is assigned a value of 0.5 to represent the fact that since the feature could not be measured nothing can be said about the QRS complex on the basis of the missing feature. Furthermore, in the extreme situation of a QRS complex on which no measurements are made all the MLP input units would be assigned a value of 0.5, from which it follows that a well trained neural network should generate an output value of 0.5. This leads to the second improvement over the standard paradigm, which is that the selection criterion for the optimum MLP is the one whose output value is 0.5 when its input units are all assigned a value of 0.5. It is demonstrated that an MLP chosen using this criterion exhibits superior classification performance compared to MLPs chosen using the standard selection criteria.

2. Methods

ECGs in which each beat has been correctly labelled as N or V were digitally filtered to eliminate or reduce low frequency baseline wander, 50 or 60 Hz mains interference, and high frequency electromyographic (EMG) noise. Then for each beat features such as the QRS width and height etc. were measured, and each feature was normalised with respect to the expected value of that feature for a supraventricular beat. A look-up table [1] was then used to convert each normalised feature, x_i, to a probability estimate, $p(N|x_i)$. This is the probability that the QRS complex is supraventricular given this feature. In this way each QRS complex is represented by a vector of single-feature probability estimates, $p(N|x_1)$, $p(N|x_2)$, …,$p(N|x_n)$. In total 440,888 vectors were generated for N beats and 42,426 vectors for V beats. This is more than ten times as many N vectors as V vectors, and using all these data to train a neural network would cause the latter to learn that 'most beats are normal'. Whilst this is true for this data set and for many ECGs in general, it is not universally true, and a neural network which has learned to bias its decisions in favour of N beats will perform poorly in some instances. To avoid this the neural network must be taught that the *a priori* probabilities that a beat is N or V are equal, i.e. $p(N) = p(V) = 0.5$, and this is achieved by means of a balanced training set containing equal numbers of N and V vectors. Such a training set was created by randomly selecting 42,426 of the 440,888 N vectors plus all of the 42,426 V vectors. A large number of neural networks were then trained to determine the following parameters:

- the best subset of features to use as inputs to the neural network (feature selection);

- the optimum neural network architecture, or more specifically how many hidden units to use since the number of input units is determined during feature selection and the number of output units is one for a two class problem;

- when to stop training.

More features, more hidden units, and longer training times enable the neural network to learn the data in its training set with greater accuracy, but risks over fitting the neural network to these data with the result that it performs poorly on new data once training is complete and the neural network is put into use. To prevent this and thereby ensure good generalisation, only half the data (i.e. 21,213 N vectors and 21,213 V vectors) were used for training and the other half were retained in a separate cross-validation set. After each pass through the training set, training was suspended and each vector in the cross-validation set was presented to the neural network's input units. The value produced at the neural network's output unit in response to each vector was compared with the desired output value, which is '1' for N vectors and '0' for V vectors. From this the m.s.e. between the desired and actual output values was calculated over the entire cross-validation set. As described earlier, the criterion for when to stop training was the number of passes through the training set that minimised the m.s.e. on the cross-validation set. By training a large number of neural networks using different subsets of input features and different numbers of hidden units an optimum neural network architecture and weights set can then be selected using either of two standard selection criteria, namely the neural network that has either the smallest m.s.e. or lowest classification error rate on the cross-validation set.

MLP neural networks with architectures of 6 to 8 input units (and therefore input features), 4 to 10 hidden units and a single output unit were trained using the error back-propagation algorithm [3] with values for the learning rate, μ, of 0.01 and momentum term, η, of 0.6. Linear activation functions were used in the MLP's input units and sigmoidal activation functions in its hidden and output units. The minimum m.s.e. on the cross-validation set was used as the criterion for when to stop training, and to account for the possibility of local minima in error space each MLP architecture was trained several times using a different random initialisation of the weights set on each occasion. This was then repeated using the cross-validation set as the training set and vice versa. A total of 720 MLPs were trained and from these the two 'optimum' MLP architectures and weights sets were identified using the standard selection criteria described earlier. A third MLP for which input values of 0.5 generate an output value of 0.5 was also identified, and the classification performance of these three MLPs was compared.

3. Results

The scatterplot in Figure 1 shows the m.s.e. on the cross-validation set (x-axis) versus the classification error rate on the cross-validation set, as a percentage, (y-axis) for each of the 720 MLPs. The arrows indicate the MLPs which have the smallest m.s.e. or lowest classification error rate on the cross-validation set (marked with an 'X'), along with the MLPs for which input values of 0.5 generate an output value of 0.5 (marked with an 'O'). The details of these four MLPs are presented in Table 1.

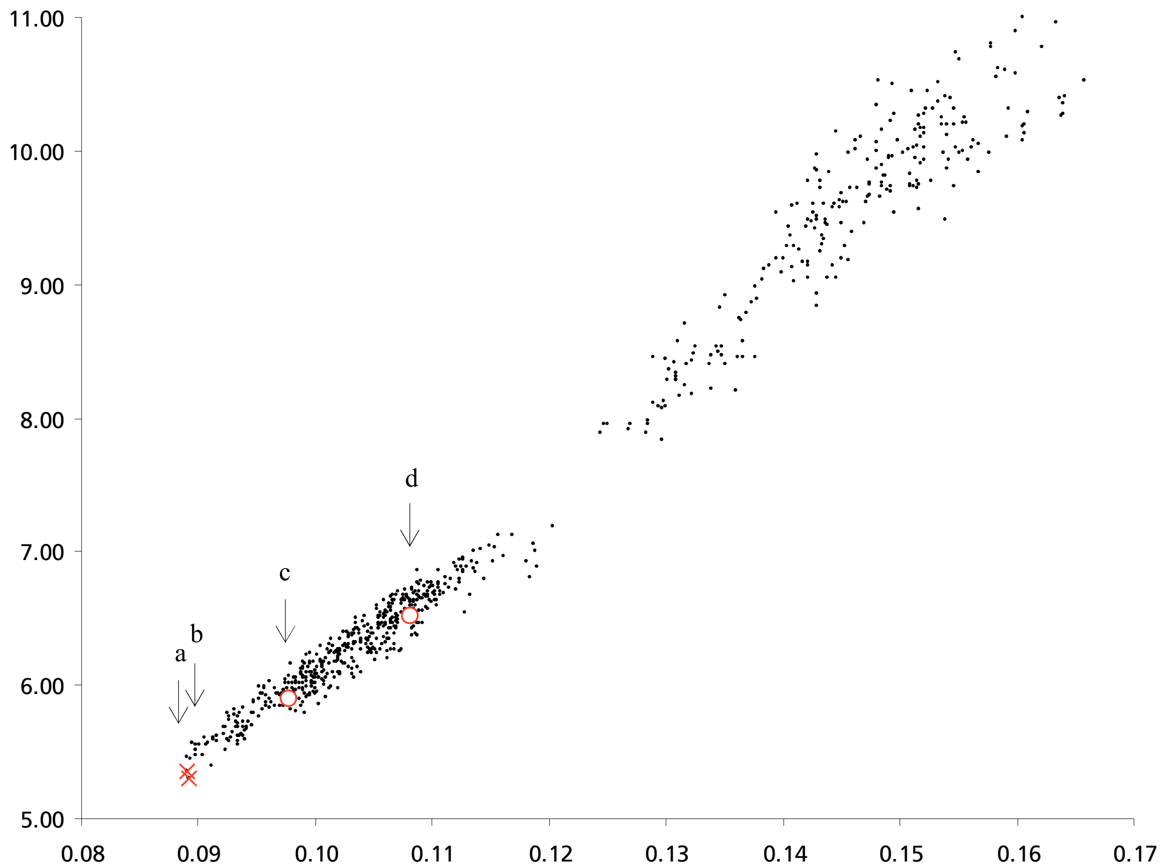

Figure 1. The m.s.e. on the cross-validation set (x-axis) versus the classification error rate on the cross-validation set, as a percentage, (y-axis) for 720 MLPs. See text for details.

Table 1. Details of the four MLPs identified in Figure 1.

MLP	architecture	m.s.e.	classification error rate
(a)	8-9-1	0.088980	5.35%
(b)	8-9-1	0.089109	5.30%
(c)	7-9-1	0.097795	5.90%
(d)	8-6-1	0.108115	6.52%

Of the two MLPs (c) and (d) for which input values of 0.5 generate an output value of 0.5, MLP (d) was selected for further comparison against MLPs (a) and (b) since all three use the same set of eight input features. These three MLPs were compared by plugging each one into Medilog ADAPT and evaluating its classification performance on an independent test set of eight 3-channel 24-hour ECG recordings from patients whose data are *not* included in the training or cross-validation sets. These ECGs contain over 770,000 QRS complexes in total, of which almost 49,000 are ventricular. 'Hands off' testing was performed in accordance with ANSI/AAMI EC38:1998 [4] and ANSI/AAMI EC57:1998 [5] and the results are presented in Tables 2–4. In these tables Nn is the number of correctly classified N beats, Nv is the number of N beats that were incorrectly classified as V, Vn is the number of V beats that were incorrectly classified as N, and Vv is the number of correctly classified V beats. The values for VEB detection sensitivity (Se) and positive predictivity (+P) are given as percentages.

The results are summarised by the gross and average statistics for VEB Se and VEB +P in Tables 2–4, which show that the performance of the MLP for which input values of 0.5 generate an output value of 0.5 is superior to that of either MLP selected using the standard criteria and uses fewer hidden units. The accuracy of the MLP in Table 4 can be calculated using the totals for Nn, Nv, Vn and Vv in conjunction with the following formula [6]:

Accuracy = (Nn+Vv) / (Nn+Nv+Vn+Vv)

which yields:

(722315+47760) / (722315+266+836+47760) = 99.9%

Table 2. Classification results for the MLP with the smallest m.s.e. on the cross-validation set.

	Nn	Nv	Vn	Vv	VEB Se	VEB +P
ECG1	60718	94	34	19704	99.62	98.73
ECG2	95386	108	135	9044	98.51	98.77
ECG3	105646	27	80	1694	95.38	93.75
ECG4	106718	10	121	2569	95.47	99.57
ECG5	111901	85	1323	11412	88.24	93.98
ECG6	68322	107	93	1838	95.09	93.39
ECG7	75933	7	47	207	81.50	95.83
ECG8	97514	0	1	148	99.33	98.67
Total	722138	438	1834	46616		
Gross					95.73	97.16
Average					94.14	96.59

Table 3. Classification results for the MLP with the lowest classification error rate on the cross-validation set.

	Nn	Nv	Vn	Vv	VEB Se	VEB +P
ECG1	60699	115	18	19720	99.70	98.63
ECG2	95385	109	131	9047	98.54	98.74
ECG3	105647	25	76	1699	95.66	95.13
ECG4	106714	14	113	2577	95.76	99.42
ECG5	111786	199	1398	11056	85.49	93.02
ECG6	68328	101	96	1836	94.98	93.58
ECG7	75937	3	53	201	79.13	98.05
ECG8	97514	0	1	148	99.33	98.01
Total	722010	566	1886	46284		
Gross					95.05	96.96
Average					93.58	96.82

Table 4. Classification results for the MLP for which input values of 0.5 generate an output value of 0.5.

	Nn	Nv	Vn	Vv	VEB Se	VEB +P
ECG1	60749	66	78	19661	99.40	98.84
ECG2	95478	16	75	9105	99.17	99.78
ECG3	105659	13	24	1750	98.54	94.44
ECG4	106710	18	32	2659	98.81	99.22
ECG5	111961	28	546	12331	95.35	94.47
ECG6	68308	121	74	1858	96.12	92.35
ECG7	75936	4	6	248	97.64	97.64
ECG8	97514	0	1	148	99.33	98.01
Total	722315	266	836	47760		
Gross					98.08	97.43
Average					98.04	96.84

4. Discussion and conclusions

The paradigm used in Medilog ADAPT for neural network beat classification has been presented and its performance evaluated. The paradigm improves on the standard paradigm for neural network classifiers in two important aspects:

First, the use of single-feature probability estimates as inputs to the neural network means that when a measurement is missing or cannot be made – for example, due to noise or artefact – the corresponding probability estimate is assigned a value of 0.5 to represent the fact that nothing is known about the QRS complex on the basis of this feature.

Second, the use of a neural network for which input values of 0.5 generate an output value of 0.5 yields superior performance compared to neural networks selected using either of the two standard selection criteria. Indeed, if the 720 MLPs in Figure 1 are ranked in ascending order using either of the standard criteria then the MLP for which input values of 0.5 generate an output value of 0.5 would be ranked 371st or 347th (i.e. 370 or 51% of the 720 MLPs have an m.s.e. < 0.108115 and 346 or 48% have a classification error rate < 6.52%). This MLP would never be selected for use despite the fact that its performance on an independent test set is superior to that of either of the MLPs ranked first. However, given its superior performance it is this neural network that is used in Medilog ADAPT (in fact three instances of it are used since beat classification is performed on each channel independently). It has been shown that using this neural network Medilog ADAPT is 99.9% accurate when tested against a database of eight 3-channel 24-hour ECG recordings. A commercial Holter system that uses the Medilog ADAPT analysis algorithm is now in use in cardiology departments worldwide [7].

References

[1] Pardey J, Jouravleva S. The next generation Holter revolution: from analyse–edit–print to analyse–print. IEEE Computers in Cardiology 2004; 31:373-6.

[2] Richard MD, Lippmann RP. Neural network classifiers estimate Bayesian *a posteriori* probabilities. Neural Computation 1991; 3:461-83.

[3] Rumelhart DE, Hinton GE, Williams RJ. Learning representations by back-propagating errors. Nature 1986; 323:533-6.

[4] ANSI/AAMI EC38:1998 Ambulatory Electrocardiographs.

[5] ANSI/AAMI EC57:1998 Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms.

[6] de Chazal P, O'Dwyer M, Reilly RB. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. IEEE Trans Biomed Eng 2004; 51:1196-206.

[7] Medilog Darwin Clinical Application Guide. Huntleigh Healthcare, 2006 (www.medilogholter.com).

Address for correspondence.

Dr. J. Pardey.
Huntleigh Healthcare, Cardiology Products Division,
Lion House, Oriental Road, Woking, Surrey GU22 8AP, UK
E-mail: james.pardey@huntleigh-diagnostics.co.uk