CASCADE-CORRELATION NEURAL NETWORKS FOR THE CLASSIFICATION OF CERVICAL CELLS

S J McKenna, I W Ricketts. A Y Cairns, K A Hussein

1. Introduction

Cervical cancer, a disease that can be prevented if detected at the pre-cancerous stage, causes the premature death of more than 2000 women in the U.K. every year. To reduce this number a national screening programme has been instituted and involves the manual inspection of cervical smears, from all sexually active females, at recommended intervals of 3 years. The labour intensive nature of the inspection process has led to considerable interest in automation.

Automating cervical cancer screening is a task which poses considerable difficulties due to the complex and heterogeneous nature of cervical smear scenes. It has been the subject of extensive research for more than 30 years and during this time various experimental systems have been developed [2]. Descriptions and comparisons of some of the most advanced systems can be found elsewhere [5].

For the past 5 years, as part of its programme of exploring the medical applications of computer vision, the Computer Vision Research Group at the University of Dundee have been researching into the automatic inspection of cervical smears. We have assembled a database containing over 2000 expertly verified cervical cell images and have investigated the performance of a number of competing techniques applied to this very demanding inspection task.

In our most recent work we examined the performance of cascade-correlation neural networks, as developed by Fahlman and Lebiere [4], to classify isolated cervical cells as either normal (benign) or abnormal (indicative of pre-cancerous changes). A set of 80 features extracted from cell images' 2D discrete Fourier transforms has been used as a basis for classification. Cascade-correlation networks have been investigated as an alternative to the commonly used back-propagation algorithm.

2. The cascade-correlation algorithm

A disadvantage with multi-layer feed-forward networks of the type usually trained using error backpropagation [7] is that the best number of hidden layers and units varies from task to task and so must be determined experimentally. If too many hidden units are used then the network will learn irrelevant details in the training set and once trained will not generalise well. Conversely, if a network is too small, it will be unable to learn the training set properly. One approach to automatically determining a good size for a network is to start with a minimal network and then add hidden units and connections as required. This is the basis of constructive algorithms.

S J McKenna, I W Ricketts and A Y Cairns can be contacted at:
MicroCentre, Department of Mathematics and Computer Science, The University,
Dundee, DD1 4HN, UK.
e-mail smckenna@uk.ac.dund.mcs

K Hussein is with the Department of Pathology, Ninewells Hospital, Dundee

Cascade-correlation is a constructive algorithm which starts life as a single-layer network to which hidden units are then added one by one. Initially, the single-layer network is trained using a method such as the delta-rule, or Fahlman's 'quickprop' algorithm which converges more quickly [3]. If this fails to achieve a sufficiently low error rate then hidden units are added until the network error becomes acceptably small.

Whenever a hidden unit is needed, a pool of candidate units, each connected to the network inputs as well as to all previously added hidden units, have their incoming weights trained. Each candidate unit is trained to maximise S, the covariance (or 'correlation') between its output value, v, and the residual network output error, δ, where p is the training pattern and $\bar{v}$ and $\bar{\delta}$ are the average values of v and δ over all the training patterns (1).

$$S = |\sum_p (v_p - \bar{v})(\delta_p - \bar{\delta})| \tag{1}$$

This maximisation is achieved by performing gradient ascent in a similar manner to the gradient descent performed by the delta rule. This is done by calculating the partial derivative of S for each incoming weight, w_i, using equation (2), where σ is the sign of the correlation between the candidate's output and the network output, f'_p is the derivative of the unit's activation function and $i_{i,p}$ is the input received from unit i. Once all the candidates have been trained, the most successful candidate is connected to the output layer and the other candidates are discarded. All the output unit weights are then trained.

$$\delta S/\delta w_i = \sum_p \sigma(\delta_p - \bar{\delta}) f'_p i_{i,p} \tag{2}$$

Initial benchmarks on artificial problems with low-dimensional input spaces suggested that this algorithm was faster to train than standard back-propagation [4]. The resulting networks had nearly as few hidden units as the best size of back-propagation network.

3. Method

A training set and a test set were each selected at random from a database of 256x256 pixel grey-level images of isolated cervical cells. Each of these sets contained 524 images, approximately 50% of which were of abnormal cells. A set of 80 texture and energy features was extracted from the frequency domain of each image. Details of this feature extraction step can be found elsewhere [1]. These 80 features were used as inputs to the neural network.

Cascade-correlation neural networks with 81 inputs (including a bias input) and a single output were used to classify cells as either normal or abnormal. Ten networks were trained using hidden units with sigmoid activation functions. A further ten were trained using hidden units with Gaussian activation functions. Pools of eight candidate units were used. Training was performed using Fahlman's quickprop algorithm. A network's training was stopped if it had not learned its training set after 16 units had been added.

One potential disadvantage with cascade-correlation is the need to determine exactly when to halt training and add a new hidden unit. In the experiments reported here, training of the weights at each stage in a network's construction was halted when learning became 'very slow' or when 500 epochs had elapsed. This strategy was chosen for convenience and for speed of operation. It is probably not the best strategy for obtaining a network which generalises well because the point at which learning becomes very slow is unlikely to correspond to the point at which the best generalisation is achieved.

4. Summary of results

Overtraining always occurred. As more and more units were added the training set error fell until, in most cases, the set was learned completely. The smallest test set error, however, was always achieved with fewer hidden units.

The networks with sigmoidal units achieved an average test set performance of 498.3(95.1%) correctly classified cells with a maximum of 501(95.6%) and a minimum of 494(94.3%). The average number of hidden units needed for generalisation was 5.8.

The networks with Gaussian units achieved an average of 496.1(94.7%) correctly classified test set cells with a maximum of 502(95.8%) and a minimum of 491(93.7%). The average number of hidden units needed for generalisation was 3.9.

5. Discussion

There seems little to choose between sigmoid and Gaussian activation functions. Pools of units containing both types might result in better performance but this has not been investigated.

An MLP with a single hidden layer of 4 units trained using backpropagation has achieved a test set classification rate of 96.3% (505 cells correct) using the same data sets [6]. Varying the number of hidden units did not improve this figure. The best generalisation achieved with cascade-correlation was a slightly lower 95.8% (502 cells correct) using 4 hidden units.

Cascade-correlation has been shown to be capable of good generalisation when applied to the problem of cervical cell classification. Experimentation with various numbers of hidden layers and units was unnecessary and training times were approximately 20 times faster than with standard backpropagation. As a result, networks exhibiting good classification accuracy were obtained quickly.

Acknowledgement

Much of the code used in this work was adapted from a public domain 'C' implementation of cascade-correlation written by R. Scott Crowder, School of Computer Science, Carnegie Mellon University, Pittsburgh.

References

[1] Banda-Gamboa H 1990 Classification of cervical cells using computer vision and the frequency domain, Ph.D. Thesis, University of Dundee

[2] Banda-Gamboa H, Ricketts I W, Cairns A Y, Hussein K, Tucker J H and Husain O A N 1992 Experimental prescreening systems for automated cervical cytology - a review, Analyt Cellular Pathology **4** 25-48

[3] Fahlman S E 1988 Faster learning variations on back-propagation: an empirical study, Proc 1988 Connectionist Models Summer School 38-51

[4] Fahlman S E and Lebiere C 1990 The Cascade-correlation learning architecture, in: Advances in neural information processing systems (ed. Touretzky D S)

[5] Data on automated cytology systems as submitted by their developers 1991 Analyt Quant Cytol Histol **13** 300-306

[6] Ricketts I W 1992 Cervical cell image inspection - a task for artificial neural networks, Network **3** 15-18

[7] Rumelhart D E, Hinton G E and Williams R J 1986 Learning internal representation by error propagation, in: Rumelhart D E, McClelland J L & the P.D.P. Research Group, Parallel Distributed Processing Vol. 1 (M.I.T. Press)