A COMPARISON OF NEURAL NETWORK ARCHITECTURES FOR CERVICAL CELL CLASSIFICATION

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

Dundee University, UK

ABSTRACT

Large-scale screening programmes are operating to reduce the incidence and mortality rate of cervical cancer, a disease which is preventable if detected at the pre-cancerous stage. Screening is based upon the manual inspection of Papanicolaou smears. This is a highly demanding and labour-intensive task and for over thirty years there has been considerable interest in automating the process.

The authors are investigating the use of various neural network architectures for the analysis and classification of smear scenes. A feature space was derived from the magnitude of the Fourier transform using a wedge-ring arrangement. The features obtained were invariant to translation and rotation. Neural nets were then used to both reduce dimensionality and to perform the classification. An expertly verified database containing over 2000 high-resolution cell images was used to measure the performance of the nets.

The single-layer perceptron, multi-layer perceptrons and the constructive algorithm of Fahlman and Lebiere were each used as classifiers. The effect of feature extraction nets for pre-processing the feature space was also investigated. Performances were compared in terms of speed, network size and ability to learn and generalise. In addition, classification by a parametric Bayesian classifier allowed comparison with a statistical method. Good classification results were obtained.

AUTOMATED PRESCREENING OF CERVICAL SMEARS

Several developed countries have introduced screening programmes in an attempt to reduce mortalities from cervical cancer. Analysis of data from some of the largest programmes indicates that screening reduces the probability of a woman developing invasive cervical cancer by approximately 90%., Eddy (1). Screening is based upon the microscopic examination of Papanicolaou smears (samples of cellular material collected from the cervix). Approximately four million smears are produced in the U.K. alone each year and a single smear contains as many as 200,000 cells. Only 5% of these smears contain any abnormal cells and many of these will have relatively few such cells. A trained cytotechnician can typically examine 50-80 smears in a day taking 5-10 minutes per smear. The task is a tedious and fatiguing one. Automatic removal of the bulk of healthy specimens from a cytology laboratory's workload would result in large savings allowing the cytologists to concentrate their efforts on the diagnosis of suspect smears.

The earliest attempts to automatically differentiate between normal and abnormal cervical cells were made in the 1950's and the possibility of automating prescreening has been researched extensively in the intervening years. An historical overview of developments can be found in Banda-Gamboa *et al.* (2). Several systems using image processing and other artificial intelligence techniques such as expert systems and neural networks are currently under development in the U.S.A. (3). In order to be accepted into cytology laboratories an automated prescreener must be fast, accurate and reliable. Neural network technology is well suited to such a task, allowing highly parallel, accurate and robust systems to be constructed.

Due to the huge amount of information contained in each smear, a dual-resolution strategy has been adopted (by both human and automated screeners). Each smear is first scanned at low resolution and areas of interest which might contain abnormal cells are then identified. These areas of interest are rescanned and analysed at an increased resolution. This paper is concerned only with the high resolution analysis of previously identified areas of interest.

THE CELL IMAGE DATABASE

Experiments were conducted using a database of greyscale images consisting of 256x256 7-bit pixels. These images were obtained from routinely prepared Papanicolaou smears using a Hitachi b/w CCTV camera mounted on a microscope fitted with a x100 oil immersion objective. An experienced cytologist located and classified the cells captured. Each image used in this study contained a single nucleus with part or all of its associated cytoplasm. Cytoplasmic material from other cells and additional artifacts

were also often present. A total of 1404 images was used. Half of these contained normal cells at various stages of maturity (superficial, intermediate and parabasal cells) while the remaining half contained cells with varying degrees of abnormality (mildly, moderately and severely dyskaryotic cells). An image set containing 50% of these images was used to train various classifiers. The remaining images were then used to test their ability to generalise.

FEATURE EXTRACTION IN THE FREQUENCY DOMAIN

After correction for shading effects (due to uneven illumination and non-uniform sensitivity of the sensing device) images were transformed to the frequency domain by applying a 2D discrete Fourier transform. This yielded a representation invariant under translation. A set of 80 features was then extracted from each frequency domain image using a ring-wedge arrangement. These features measured energy and texture of the frequency domain image. Details of this feature extraction process may be found elsewhere, Banda-Gamboa (4). Classification of images as either normal or abnormal was subsequently based upon these 80 features.

An apparent advantage to this method is that no high-resolution segmentation of the image is used in order to extract the features. High-resolution segmentation of cells has been the most difficult and the most important step in cervical smear screening systems with inaccurate segmentation leading to the extraction of erroneous features. In contrast, the features extracted here are not dependent upon accurate segmentation.

CLASSIFICATION

Initially, a parametric Bayesian classifier was used to classify cell images as either normal or abnormal. Subsequently, it was decided to compare it with various neural network classifiers with the dual aims of improving classification accuracy and providing a benchmarking study for the different neural network architectures used.

Parametric Bayes classifier

This classifier is based upon Bayes' rule which assigns an object to the class with the highest conditional probability, Duda and Hart (5). It calculates a linear discriminant function so as to minimise the number of misclassifications under the assumptions that the feature vector is drawn from a multivariate normal distribution and that all classes have identical covariance matrices.

Back-propagation classifiers

Fully-connected feedforward neural networks were trained using error back-propagation, Rumelhart et al. (6). Both single layer networks (SLPs) and multi-layer networks with one or two hidden layers (MLPs) were used. Network units had symmetric sigmoid activation functions with range (-0.5,0.5). All input and output patterns were scaled in the range [-0.5,0.5]. A value of 0.1 was added to the derivative of each unit's activation function in order to avoid derivatives of zero ('flat spots'), Fahlman (7). These alterations were found to significantly decrease training times. A learning rate of 0.003 and a momentum term of 0.9 were used. These parameters were set in accordance with the rule of Eaton and Olivier (8) which was found to result in good convergence.

Cascade-correlation

The cascade-correlation algorithm of Fahlman and Lebiere (9) is a constructive algorithm which starts life as a single layer network to which hidden units are added one by one until a sufficiently low error is achieved. It attempts to automatically generate a network with a suitable topology and thus avoid the need for lengthy experimentaion with different numbers of hidden units and layers usually associated with the use of multi-layer networks. All training is performed using Fahlman's quickprop method.

Initial benchmarks suggested that cascade-correlation was considerably faster to train than standard back-propagation and resulted in networks with nearly as few hidden units as the best size of back-propagation network found (9). In a series of experiments on real-world pattern classification tasks reported in Yang and Honavar (10), cascade-correlation was found to learn faster than back-propagation but did not generalise as well on two out of the three data sets studied. It did, however, generalise better than back-propagation on the third data set for which no hidden units were needed.

Cascade-correlation networks with sigmoid activation functions were trained. Pools of 8 candidate units were used. An offset of 0.1 was added to the derivative of each unit's activation function in order to avoid flat spots. After experimentation with different parameter values the learning rate was set to 0.75, input weight decay to 0.0, output weight decay to 0.0001, patience to 8 and weight change thresholds to 0.001. A maximum of 10 hidden units were allowed to be added.

TABLE 1 - Comparison of network performance (Values in parentheses indicate standard deviations)

	# weights	Avg. epochs to train	Avg. test set error(%)	Min. test set error(%)
SLP (80-1)	81	1745 (54)	10.6 (0.1)	10.4
Cascade-corr.	502 (205)	1963 (550)	11.5 (1.0)	10.1
Bayesian	-	-	8.7	8.7
MLP (80-6-2-1)	503	2180 (1601)	7.3 (0.4)	6.7
MLP (80-6-1)	493	1965 (1051)	7.1 (0.3)	6.6

PERFORMANCE COMPARISONS

Each of the network architectures was trained 10 times with different initial random weights in the interval [-0.3,0.3]. In the case of MLPs with one hidden layer, 6 units was found, after experimentation, to be a good size for the hidden layer. MLPs with a second hidden layer of 2 units were also trained.

Table 1 compares the performance of the various classifiers in terms of the number of connection weights, the number of iterations through the training set (epochs) required and most importantly their ability to generalise on the test set. Classification accuracy was measured in terms of the percentage of test set images correctly classified. The values given in parentheses indicate standard deviations. It should be noted that the set of 80 features used was arrived at after much experimentation with the Bayesian linear discriminant classifier. The features have been specially tailored to suit this particular classifier.

The SLP and cascade-correlation networks both yielded higher test set errors than the Bayesian classifier. Unexpectedly, cascade-correlation did not always outperform SLP. This was largely due to the mechanism used by cascade-correlation to decide when to suspend training and add a new hidden unit. The point at which learning slowed below the rate determined by the 'patience' and 'weight change threshold' parameters varied from trial to trial so that even with careful selection of these parameters it was unlikely that the point selected would be the best at which to halt and add a new unit.

MLP networks were consistently more accurate than the Bayesian classifier. The improvement in test set performance was statistically significant ($P<0.001$). The lowest test set error, obtained using a single hidden layer of 6 units, was 6.6%. Only MLP networks learned to classify the entire training set correctly.

DIMENSIONALITY REDUCTION USING NEURAL NETWORKS

Principal components analysis (5) of the feature vectors showed there to be many small eigenvalues. This indicated that there was a lot of redundancy in the data suggesting that its dimensionality could be usefully reduced. A widely used technique for dimensionalilty reduction is to project the data onto its principal components, and a similar result can be achieved with neural networks.

Sanger proposed a rule to train an unsupervised single-layer network of M units so that its weight vectors converged to the first M eigenvectors of the autocorrelation matrix of its inputs, Sanger (11). (These are equal to the first M principal components when zero-mean data is used). Oja proposed a similar rule which makes the weight vectors converge to span the same subspace as the first M eigenvectors, Hertz (12). In Oja's rule, instead of each unit finding a particular eigenvector, the units form a distributed representation with equal variance at each neuron. Other authors have suggested similar rules and the reader is refered to Oja (13) for references to these.

Hrycej improved the generalisation ability of a back-propagation classifier by preprocessing its input data with this type of network, Hrycej (14). A rule which finds a distributed representation like that of Oja was found to yield a better classification rate than a rule which found particular eigenvalues.

An N-input linear self-supervised (or 'auto-associative') back-propagation network (SSBP) with one

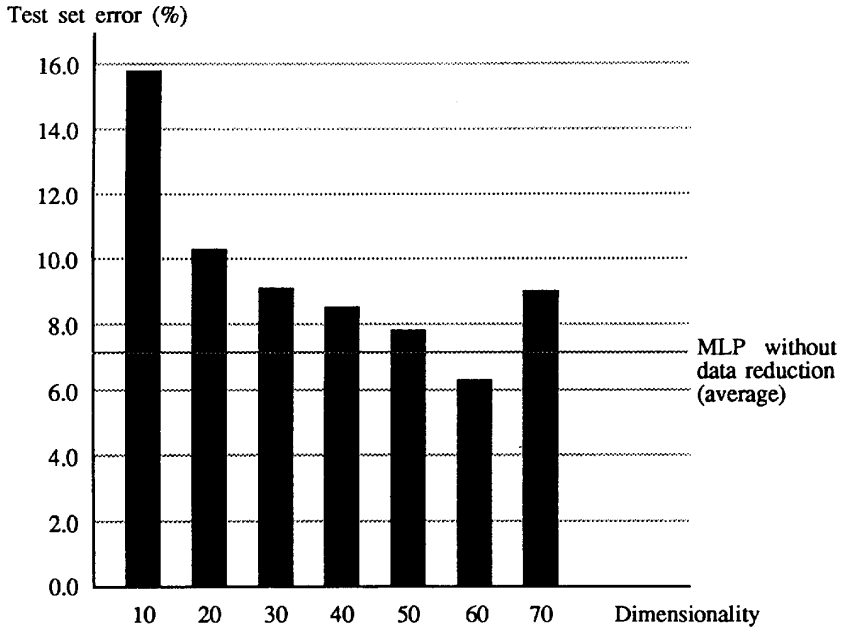

Figure 1: The effect of SSBP hidden layer size upon MLP classifier generalisation

hidden layer of M<N units forms a distributed representation at its hidden layer which, like the result of Oja's rule, is a linear combination of the first M eigenvectors, Baldi and Hornik (15).

DIMENSIONALITY REDUCTION AND CLASSIFICATION BY BACKPROPAGATION

SSBP networks were used to reduce the dimensionality of the 80-element feature vectors. The outputs of the hidden units were used as inputs to an MLP classifier. Although its convergence properties are not as good as those of the previously described single-layer networks, SSBP was chosen to perform the dimensionality reduction in order to maintain homogeneity. Both networks could then be trained by the same method, namely back-propagation.

Figure 1 shows the test set misclassification rates obtained by reducing the data dimensionality from 80 to 10, 20, 30, 40, 50, 60, and 70. Preprocessing the data with a 60 hidden unit SSBP resulted in a slightly reduced test set error. An MLP with 60 inputs and a single hidden layer of 3 units obtained a test set misclassification rate of 6.3%. Dimensionalities 51-59 and 61-69 were not used in this study and it is possible that they could lead to even lower error rates.

SUMMARY

An 80-element feature set extracted from the frequency domain was used as the basis for classification of cervical cell images. The features were extracted without the need for accurate segmentation of the cells. Classifiers used were SLP, MLP and cascade-correlation networks as well as a non-neural Bayesian classifier.

The test set misclassification rates obtained using SLP and cascade-correlation networks were higher than that of the Bayesian classifier. MLPs, however, were able to consistently outperform the Bayesian classifier.

An SSBP network was used to form a reduced representation of the feature data at its hidden units. This representation was then used as input to an MLP classifier. This scheme resulted in a slight decrease in test set misclassifications. Further experiments are needed to explore the full potential of this approach.

FURTHER DEVELOPMENTS

In assessing the utility of these results for automated prescreening it is useful to consider the errors made by laboratories using manual screening. Approximately 60% of screening errors occur because

abnormal cells, though present in the cervix, do not appear on the smear. A further 40% of screening errors occur because abnormal cells are missed by the cytologist. Errors due to cells being examined and incorrectly classified are very rare, Wilkinson (16). Therefore the accuracy of the best classifier trained here is not good as that of human screeners.

In order to improve performance, experiments are being carried out with images of a higher quality. These images have increased spatial resolution (512x512 pixels) and intensity resolution (8 bits). In preliminary experiments using a small database of 318 images a backpropagation network was able to classify its test set with 100% accuracy. This new image database is currently being expanded.

ACKNOWLEDGEMENTS

The cascade-correlation experiments were conducted using C code written by R. Scott Crowder of Carnegie Mellon University (with some minor modifications). The authors wish to thank Dr. Fahlman and Mr. Crowder for making this code available.

REFERENCES

1. Eddy D. M., 1990, Annals of internal medicine, 113, 214-225

2. Banda-Gamboa H., Ricketts I. W., Cairns A. Y., Hussein K. A., Husain O. A. N., 1992, Analyt Cell Path, 4, 25-48

3. Data on automated cytology systems as submitted by their developers1991, Analyt Quant Cytol Histol, 13, 300-306

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

5. Duda R. O. and Hart P. E., 1973, Pattern classification and scene analysis, John Wiley & Sons

6. Rumelhart D. E., Hinton G. E., Williams R. J., 1986, Learning internal representation by error propagation, in: Parallel distributed processing vol. 1, Rumelhart D. E. and McClelland J. L. (eds.), M.I.T. Press, Cmbridge, MA.

7. Fahlman S. E., 1988, Faster learning variations on back-propagation: an empirical study, Proc. 1988 connectionist models summer school, 38-51

8. Eaton H. A. C. and Olivier T. L., 1992, Neural networks, 5, 283-288

9. Fahlman S. E. and Lebiere C., 1990, The cascade-correlation learning architecture in: Advances in neural information processing systems, Touretzky D. S. (ed.), 524-532

10. Yang J. and Honavar V., 1991, Experiments with the cascade-correlation algorithm, I.J.C.N.N. vol 3., 2428-2433

11. Sanger T. D., 1989, Neural Networks, 2, 459-473

12. Hertz J., Krogh A. and Palmer R. G., 1991, Introduction to the theory of neural computation, Addison Wesley

13. Oja E., 1992, Neural Networks, 5, 927-935

14. Hrycej T., 1992, Neurocomputing, 4, 17-30

15. Baldi P. and Hornik K., 1989, Neural Networks, 2, 53-58

16. Wilkinson E. J., 1990, Clinical Obstetrics and Gynecology, 33, 817-825