Quantification of Diabetic Retinopathy using Neural Networks and Sensitivity Analysis

Andrew Hunter, James Lowell, Jonathan Owens, Lee Kennedy
School of Computing and Engineering Technology, University of Sunderland
Sunderland, Tyne and Wear, UK.
Andrew.Hunter@sunderland.ac.uk

David Steele
Sunderland Royal Eye Infirmary
Sunderland, Tyne and Wear, UK.

Abstract

The design of neural network classifiers for the identification of diabetic retinopathy is discussed. Red-free digitised fundal images are tiled, and a neural network is trained to distinguish exudates from drusen (similar appearing lesions). By quantifying the degree of retinopathy, the approach can be used to screen diabetic patients for referral. A novel form of hierarchical feature selection using sensitivity analysis is presented. The resulting neural network is compact, and achieves 91% sensitivity and specificity on a test set.

1 Introduction

Diabetic retinopathy is the commonest cause of blindness in the working age group within developed countries [1]. Early detection and treatment allows a 50% reduction in the incidence of visual loss [2]. A comprehensive screening program needs to process 30,000 diabetes mellitus patients per million population. However, the grading of fundal images is time-consuming and requires highly skilled staff.

This paper discusses the development of an automated screening tool. The system takes a quantitative approach, identifying the proportion of the fundal image that contains exudates. This can identify new cases, monitor the progress of existing cases, and grade the stage of retinopathy. Rather than identifying individual exudates (which vary significantly) the system tiles the image, and determines whether each tile contains exudates. Hence a very high level of sensitivity and specificity per tile is not necessary, as the effects of misclassifications of individual tiles tend to cancel each other out, in the statistical sense.

A number of related systems have been reported in the literature. They each attempt to detect and quantify features in the fundal image that are characteristic of diabetic retinopathy, typically microaneurysms and exudates. Neovascularisation is also characteristic of advanced retinopathy, but we are not aware of any systems that check for this feature.

Microaneurysms are the earliest indicator, and a system that counts micro-

aneurysms has been suggested by Phillips and Spencer [3][4]. However, this requires a fluorescein angiogram, and so is not suitable for large-scale screening.

Exudates can be identified by thresholding preceded by shade correction [5]. A drawback is that other lesions (drusen, cotton wool spots and laser scars) can be mistakenly identified. Goldbaum has suggested using the lesion colour to distinguish exudates [6]; however, this approach is not particularly robust. Gardner [7] has suggested using neural networks, with one network to distinguish exudates from each of the other lesion types. After median smoothing, the red-free values were fed directly into a large neural network (using 20x20 tiles, the network had 400 inputs).

In this paper we propose a novel approach to feature selection that radically reduces the complexity of the neural network, thus increasing performance, reliability, training and execution speed. The method is hierarchical, with individual feature selection stages feeding small feature subsets through to further stages. Each stage involves training a neural network, and analysing the contribution of features using a recently introduced form of sensitivity analysis [8]. Since each individual network is reasonably compact, and sensitivity analysis is very fast, the method can be used to select from among hundreds of features extremely quickly.

The final network designed by this method has only eleven inputs and seven hidden units. It achieved 91% correct classification on the test set, compared with 60% for a network trained directly on the red-free image.

In section 2 we describe sensitivity analysis, and its application in feature selection. In section 3 we report on our experiments, and describe our approach to hierarchical feature selection. Section 4 concludes the paper.

2 Sensitivity Analysis

In sensitivity analysis, a network is analysed and an index of importance assigned to each input. The most common approach is to explicitly perturb each input in turn [9], or to calculate the gradient of the output with respect to each input [10], (inputs that have little effect on the output when changed are unimportant). We have recently introduced an alternative method that avoids some drawbacks of the aforementioned [8]. In neural network modelling, if an input is unavailable, a missing value substitution procedure is typically used (e.g. substitution of mean). For each input in turn, we calculate the sensitivity as $s_i = e_i/e$, where e_i is the network error with mean substitution of input i, and e is the normal error.

Sensitivity analysis indicates the significance of input variables in a particular neural network. This can in turn be used to select features for use in further modelling, although caution is necessary. Variables may demonstrate both inter-dependence and mutual redundancy [11], one consequence of which is that a number of networks trained on the same data may produce different sensitivity rankings. However, experience shows that the most significant few variables are consistently identified in the majority of cases. A single training run is usually sufficient to identify key variables, as the sensitivity of variables calculated using the training and test sets can be compared for consistency. Repeated training runs can be used for increased confidence.

3 Experimental Results

We divided the red-free fundal images (simulated by using the green component) into 16x16 tiles. Fast Fourier transforms and Prewitt edge-detection filters were applied. This provided three basic feature sets (*red-free*, *fourier* and *prewitt*). The basic feature sets were each processed using four techniques: summary statistics (Mean, Standard Deviation, Skew and Kurtosis), first 16 principal components, 16-bin histogram, and principal components of the histogram.

Neural networks were built using each of the twelve methods of processing, then key features selected using sensitivity analysis. The key features were combined into three composite feature vectors, and the process repeated to determine a single feature vector. Another pass of sensitivity analysis established a feature vector for the final model.

The data were taken from sixteen images captured by a conventional fundal imaging camera, and digitised from slides. Tiles were classified by an ophthalmologist using custom software; a range of representative tiles was selected, including some boundary cases only partially occupied by the lesion. Image processing (fourier and prewitt transforms) was performed using WiT image processing software; the raw output images were converted to data files by custom software. Neural models, sensitivity analysis, ROC curve and classification results were performed using the Trajan neural network package.

The data set contained 95 tiles featuring drusen, and 116 featuring exudates. The relatively small number of cases adds to the difficulty of the feature selection process. The data was divided into a training set (44 drusen, 56 exudate) and a test set (51 drusen, 60 exudate).

Standard MLP networks were used (logistic activation function), with a small number of hidden units to reflect the low data volume available (between 5 and 8, depending on the number of inputs). Input variables were minimax normalised. Weigend weight regularisation [12] was deployed, with $\lambda = 0.005$ to help restrict model complexity. Weights were randomly initialised in the range [-1,+1].

Training was by 300 epochs of on-line back propagation (learning rate 0.1, momentum 0.3), followed by 300 epochs of conjugate gradient descent. This is an efficient combination with small data sets - if CGD alone is used, it tends to exhibit over-learning very rapidly, with training and test errors diverging rapidly. However, once BP has identified a good minimum, CGD improves terminal convergence and is much more likely to remain stable. Regularisation was extremely important - without it, over-learning invariably occurred and results were inferior.

ROC curves [13] were generated to compare results. The classification threshold was set to equalize sensitivity and specificity. Performance was assessed using the proportion of test cases correctly classified at this threshold. This statistic is highly correlated with the area under the ROC curve.

As a benchmark, we trained neural networks with 256 inputs (the red-free pixel values). Gardner et. al. [7] reported over 90% performance with a 400 input network (using 20x20 tiles) with 80 hidden units. However, they used 1,000 training cases, whereas we were limited to 100; consequently, we encountered severe over-learning

when building such large networks, and performance of only 52% was achieved. With the number of hidden units reduced to five, and using weigend regularisation, 60% performance was achievable. With more training data we could probably raise this, but not to the levels reported by Gardner et. al. The major cause of the disparity is no doubt a difference in the inherent difficulty presented by the data sets.

Table 1 summarises the results of the hierarchical feature selection procedure. The number of inputs to each network, and the number of these selected by sensitivity analysis, is shown together with the network's performance. The *Red-free*, *Fourier* and *Prewitt* composite networks use the features identified by the corresponding *Summary*, *PCA*, *Histogram* and *PCA Histogram* networks. The *Overall Composite* network use the features identified by the individual composite networks; sensitivity analysis on it removes a further four variables. Thus, eleven variables are used in training the final model.

In general the test rates follow the expected pattern, with higher values appearing as more significant variables with different characteristics are united in progressively higher composite levels. A minor exception is seen in the prewitt composite nework, which has slightly lower performance than the prewitt summary statistics network. This is due to the marginal value of the non-summary features in the prewitt set, combined with the inherent variability in the training process.

Table 1: Hierarchical feature selection by sensitivity analysis

Model		*Inputs*	*Selected*	*Test Rate*
Red-free	Summary	4	3	0.71
	PCA	16	3	0.62
	Histogram	16	3	0.63
	PCA Histogram	16	4	0.64
	Composite	13	6	0.83
Fourier	Summary	4	3	0.73
	PCA	16	4	0.62
	Histogram	16	1	0.67
	PCA Histogram	16	2	0.59
	Composite	10	4	0.77
Prewitt	Summary	4	2	0.77
	PCA	16	1	0.73
	Histogram	16	4	0.53
	PCA Histogram	16	4	0.73
	Composite	11	5	0.76
Overall Composite		15	11	0.89
Final Model		11	-	0.90

The key stage is the selection of variables using the sensitivity analysis. This is done by eye, although it could be automated. Table 2 shows the results of one sensitivity analysis (fourier summary statistics). The sensitivity ratios are shown for both training and test sets. Kurtosis is identified as having extremely low sensitivity

for both sets, and can be eliminated (a sensitivity of 1.0 indicates that a feature is entirely ignored, with progressively larger figures indicating greater significance). The sensitivities of the other three variables are low but significant, and the ranking inconsistent, so they are retained. In later processing, the composite fourier network eliminates Mean and Skew, but S.D. is retained right into the final model.

Table 2: Sensitivity analysis of fourier summary statistcs

	Mean	S.D.	Kurtosis	Skew
Training	1.067	1.034	1.010	1.092
Test	1.049	1.088	1.001	1.053

The final neural network identified has 11 input variables. This neural network has 91% performance. Sensitivity analysis confirms that all the inputs are significant. The area under the ROC curve is 96.4%.

An interesting observation is that in several of the PCA networks, the first principal component has extremely *low* sensitivity, with the second and perhaps some following components highly significant. This may indicate that the first principal component corresponds to "uninteresting" or "obvious" variation, whereas the succeeding ones capture real structure.

4 Conclusion

We have demonstrated that it is possible to build a screening system for diabetic retinopathy, using neural networks and standard image processing to distinguish a key indicator of retinopathy (exudates) from other lesions. The approach is quantitative, assessing the area of the retina containing exudates. The method achieves a high level of performance (91% correct test set classification rate).

We have introduced a hierarchical feature selection method based on sensitivity analysis, which allows key features to be selected from a very large set with minimal computational effort, and despite having a small data set. The technique is well-suited to image processing domains, where a large number of candidate features can easily be generated, but determining which to use is difficult.

4.1 Future Work

The initial study has been conducted using a relatively small number of images. Our first requirement is therefore to repeat the experiments using a larger data set.

We also need to design the other parts of the classication system: lesion detection, and discrimination of exudates from cotton wool spots and laser scars.

Alternative features will be considered, including some that are sensitive to "blobs", such as Gaussian kernel convolutions and Gabor transforms, and colour information. Our approach is of course well-suited to investigating and integrating additional features with minimal effort.

References

[1] Williams, R. Diabetes mellitus. In: Stevens. A., Raftery, J. (eds). Health Care Needs Assessment. Oxford University Press, Oxford, 1994, pp. 31-57.

[2] Singer, D.E., Nathan, D.M., Fogel, H.A. Schachat, A.P. Screening for diabetic retinopathy. Ann. Intern. Med. 1992; 116: 660-671.

[3] R.P. Phillips, P.G. Ross, M. Tyska, P.F. Sharp and J.V. Forrester. Detection and quantification of hyperfluourescent leakage by computer analysis of fundus fluorescein angiograms. Graefe's Arch Clin Exp Ophthalmol 1991; 229: 329-335.

[4] T. Spencer, J.A. Olson, K.C. McHardy, P.F. Sharp and J.V. Forrester. An Image-Processing Strategy for the Segmentation and Quantification of Microaneurysms in Fluorescein Angiograms of the Ocular Fundus. Computers and Biomedical Research 1996; 29: 284-302.

[5] R.P. Phillips, J. Forester and P. Sharp. Automated detection and quantification of retinal exudates. Graefe's Arch Ophthalmol 1993; 231: 90-94.

[6] M.H. Goldbaum, N.P. Katz, M.R. Nelson, L.R. Haff. The discrimination of Similarly Colored Objects in Computer Images of the Ocular Fundus. Investigative Ophthalmology and Visual Science 1990; 31 (4): 617-623.

[7] Gardner, G.G., Keating, D. Williamson, T.H. and Elliot, A.T. Automatic detection of diabetic retinopathy using an artificial neural network: a screening tool. Br J. Ophthalmol. 1996; 80: 940-944.

[8] Hunter, A. Application of Neural Networks and Sensitivity Analysis to improved prediction of Trauma Survival. Computer Methods and Algorithms in Biomedicine (in press).

[9] T.D. Gedeon. Data Mining of Inputs: Analysing Magnitude and Functional Measures. Int. Journal of Neural Systems 1997; 8 (2): 209-218.

[10] J.M. Zurada, A. Malinowski and I. Cloete, Sensitivity Analysis for Minimization of Input Data Dimension for Feedforward Neural Network, IEEE International Symposium on Circuits and Systems, London, May 30-June 3, 1994.

[11] A. Jain and D. Zongker Feature Selection: Evaluation, Application and Small Sample Performance. IEEE Trans. Pattern Analysis and Machine Intelligence 1997; 19 (2).

[12] Weigend, A.S., Rumelhart, D.E. and Huberman, B.A. Generalization by weight-elimination with application to forecasting. In: Lippmann, R.P., Moody, J.E. and Touretsky, D.S. (eds). Advances in Neural Information Processing Systems 1991; 3: 875-882. San Mateo, CA: Morgan Kaufmann.

[13] M.H. Zweig and G. Campbell. Receiver-Operating Characteristic (ROC) Plots: A Fundamental Evaluation Tool in Clinical Medicine. Clin. Chem 1993; 39 (4): 561-577.