Improved Neural Network-based Interpretation of Colonoscopy Images Through On-line Learning and Evolution

George D. Magoulas, Vassilis P. Plagianakos* and Michael N. Vrahatis*
Department of Information Systems and Computing, Brunel University,
Uxbridge UB8 3PH, United Kingdom
Phone: +44-1895-274000, Fax: +44-1895-251686
email:George.Magoulas@brunel.ac.uk
* Department of Mathematics, University of Patras, 26110, Patras, Greece
Phone: +30-61-997374, Fax: ++30 61 992965
email:{vpp,vrahatis}@math.upatras.gr

ABSTRACT: In this work we explore on-line training of neural networks for interpreting colonoscopy images through tracking the changing location of an approximate solution of a pattern-based, and, thus, dynamically changing, error function. We have developed a memory-based adaptation of the learning rate for the on-line Backpropagation (BP) and we investigate the use of this scheme in an on-line evolution process that applies an on-line BP-seeded Differential Evolution Strategy to (re-)adapt the neural network to modified environmental conditions. We compare this hybrid strategy to other standard training methods that have traditionally been used for training neural networks off-line. Preliminary results in interpreting colonoscopy images and frames of video sequences suggest that networks trained with this strategy detect malignant regions of interest with high accuracy. Extensive testing in interpreting more complex regions is necessary to fully investigate the properties, the effect of the heuristic parameters and the performance of the hybrid learning strategy in this context.

KEYWORDS: Minimally invasive imaging procedures, Backpropagation networks, Medical image interpretation, On-line learning, Differential Evolution Strategies, Artificial evolution.

INTRODUCTION

The need for more effective methods of early detection of cancer - such as those that computer-assisted medical diagnosis systems aim to provide is obvious. In technical terms, the problem in automatic image interpretation is to associate sets of pixels (structures) in an image with the unknown objects that are present in the scene from which the image has been drawn. The difficulty increases when several objects of different kinds, related by a set of spatial-temporal relations, are present in the observed scene. In medical practice, endoscopic approaches and other minimally invasive techniques (for example, computed tomography, ultrasonography, confocal microscopy, computed radiography, or magnetic resonance imaging) are now permitting visualization of previously inaccessible regions of the body. Their objective is to increase the expert's ability in identifying malignant regions and decrease the need for intervention while maintaining the ability for accurate diagnosis. Furthermore, it may be possible to examine a larger area, studying living tissue in vivo - possibly at a distance [1] - and, thus, minimise the shortcomings of biopsies, such as a limited number of tissue samples, a delay in diagnosis, and discomfort for the patient. In diagnostic endoscopy, the medical expert, based on a distributed percept of local changes, interprets the physical surface properties of the tissue - such as the roughness or the smoothness, the regularity, and the shape - to detect abnormalities. Adjacent surfaces showing different surface properties are distinguished on the basis of the texture differences. It is important to note, however, the vast difficulties in physical attributes of the organs. For example, in colonoscopy, no two colons are alike. Even within the same colon, one section may have very different characteristics from another. This fact introduces severe limitations in the use of computer-assisted endoscopy for interpreting colonoscopy images [2]. Given a medical image, the "true" features associated with the physical surface properties of the tissue are not exactly known to the image-interpretation system developer. Usually, one or more feature-extraction models [3] are used to provide values for each feature's parameters. The findings are then used to infer the correct interpretation. On this same task of interpretation on the basis of local changes on the properties of the tissue under examination, the performance of human perception is considered outstanding. Furthermore, medical experts have the ability to either add or remove components from an image to give meaning to

what they see. Medical experts can also adapt to changes to the extent that even a distorted image can be recognized. Neural network-based methodologies present some human-like qualities, such as learning from experience, generalisation, and handling uncertainty and ambiguity in distorted or noisy images. Thus, such methods provide human experts with significant assistance in medical diagnosis [4],[5],[6],[7].

In this work we will focus on neural network-assisted endoscopy (a narrow pipe like structure, an endoscope, is passed into the patient's body) for interpreting colonoscopy images. Video endoscopes have small cameras in their tips, when passed into a body, what the camera observes is displayed on a television monitor (see Figure 1 for some of the frames of a video sequence). The physician controls the endoscope's direction using wheels and buttons and the whole procedure is carried out under variable perceptual conditions (shadings, shadows, lighting condition variations, reflections etc.).

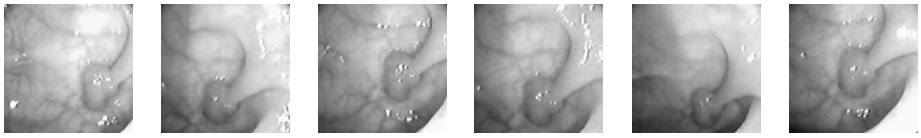

Figure 1: Six of the frames of a video sequence showing a polypoid tumor of the colon.

The use of neural networks for detecting malignant regions in these video sequences encounters several problems due: to the time varying nature of the process, to changes in the perceptual direction of the physician, and to variations in the diffused light conditions. In most of these cases, the training set is not able to represent all possible variations of the environment in which the neural network is to be operated. On-line training and retraining are suggested as possible alternatives because they allow the network to update its weights during operation by taking into account both the already stored knowledge and the knowledge extracted from the current data. Of course, the main challenge when dealing with this approach is to balance the information related to recently acquired data with the information already embodied in the network [8],[9].

In this paper, we explore on-line training and retraining of neural networks for detecting malignant regions in colonoscopy images though a formulation of the problem that is based on the idea of tracking the moving "optimum" of a dynamically changing pattern-based error measure. This approach coincides with the way adaptation on the evolutionary time scale is considered [10], and allows us to explore and expand further research on the tracking performance of evolution strategies and genetic algorithms [10],[11],[12]. However, the reader should keep in mind that in this paper we do not seek global minimisers of the error function, but we are interested in developing an on-line evolution strategy that will converge to an approximation of the optimum solution (the interesting topic of finding global minimisers in neural networks training is described elsewhere [13]).

The paper is organised as follows: the next section describes the on-line evolution strategy. Then we present some experimental results and discuss about the findings.

ON-LINE EVOLUTION STRATEGY

On-line training in neural networks is related to updating the network parameters after the presentation of each training example, which may be sampled with or without repetition. On-line training may be the appropriate choice for learning a task either because of the very large (or even redundant) training set, or because of the slowly time-varying nature of the task. Although batch training seems faster for small-size training sets and networks, on-line training is probably more efficient for large training sets and networks. It helps escaping local minima and provides a more natural approach to learning in non-stationary environments. On-line methods seem to be more robust than batch methods as errors, omissions or redundant data in the training set can be corrected or ejected during the training phase. Additionally, training data can often be generated easily and in great quantities when the system is in operation, whereas they are usually scarce and precious before. Lastly, on-line training is necessary in order to learn and track time varying functions and continuously (re-)adapt in a changing environment.

Despite the abundance of methods for learning from examples, there are only few that can be used effectively for on-line learning. For example, the classic batch training algorithms cannot straightforwardly handle nonstationary data. Even when some of them are used in on-line training there exists the problem of "catastrophic interference", in which training on new examples interferes excessively with previously learned examples leading to saturation and slow convergence [14]. Below we present an on-line BP-seeded *Differential Evolution* (DE) strategy for on-line neural network training. Firstly, we briefly present the on-line BP learning stage of the proposed strategy. We, then, proceed

by describing the on-line DE stage. Note that the description below focuses on the problem of adapting the weights on-line, assuming that the DE is always activated and does not require the input and desired output data to be known a priori. Our experiments, reported in the next section, were also conducted under the same assumptions (note, however, that in practice, whenever the changes of the environment are not considered significant and the performance is satisfactory, the weights and structure of the network should remain the same).

ON-LINE BACKPROPAGATION LEARNING

On-line BP schemes are usually based on the use of stochastic gradient descent due to the inherent efficiency of this method in time-varying environments [14],[15],[16],[17],[18]. However, sensitivity to learning parameters is a common drawback of these schemes [19]. Note that in this context, it is not possible to use advanced optimisation methods, such as conjugate gradient, variable metric, simulated annealing etc., as these methods rely on a fixed error surface [19]. In [20], a variant of the on-line BP has been proposed which exhibits improved performance when compared with other methods of the same type. A key point of the method is the use of a new learning rate adaptation schedule that exploits gradient related information from the current as well as the two previous pattern presentations:

$$\eta^{k+1} = \eta^k + \gamma_1 \left\langle \nabla E_{p-1}(w^{k-1}), \nabla E_p(w^k) \right\rangle + \gamma_2 \left\langle \nabla E_{p-2}(w^{k-2}), \nabla E_{p-1}(w^{k-1}) \right\rangle . \quad (1)$$

In (1), $\langle .,. \rangle$ stands for the usual inner product in $\Re^n$, E_p is the pattern-based error measure and ∇E_p is the corresponding gradient vector; η is the learning rate, and γ_1, γ_2 are the meta-learning rates. At the start of the learning procedure, $k = 0$, the learning rate is set to a small positive value. Then, the weights are updated on-line, for each pattern *p*, following the iterative scheme:

$$w^{k+1} = w^k - \eta^k \nabla E_p(w^k) . \quad (2)$$

The behaviour of this method in the simulations was characterised by increased speed and a higher possibility of good performance when compared with the on-line BP schemes proposed by Almeida et al. [15], and the classic on-line BP. In our experiments, we have found that the above on-line BP scheme is particularly efficient at finding a good initial approximation of the solution and we use it to initialise the population of the DE strategy.

DIFFERENTIAL EVOLUTION OF LEARNING

Evolution Strategies (ESs) are adaptive stochastic search methods that mimic the metaphor of natural biological evolution. The main differences between ESs and Genetic Algorithms lie in that the self-adaptation of the mutation operator is a key feature of the ESs, and in that GAs prefer smaller mutation probability (rate) [10],[11]. Here we use the Differential Evolution strategies, which have been designed as stochastic parallel direct search methods that can handle non-differentiable, non-linear and multimodal objective functions efficiently, and require few easily chosen control parameters [21]. Experimental results have shown that DE strategies have good convergence properties and outperform other evolutionary algorithms [21]. To apply DE strategies to neural network training we start with a specific number (*NP*) of *n*-dimensional weight vectors, as initial population, and evolve them over time; *NP* is fixed throughout the training process and the weight population is initialised by perturbing the approximate solution provided by the on-line BP. In this case, the on-line BP seeds the DE, i.e. a preliminary solution is available by the on-line BP, so the initial population might be generated by adding normally distributed random deviations to the nominal solution. However, in the experiments reported in the next section we have used a uniform distribution to perturb the approximated solution provided by on-line BP to test the robustness of our approach to noise.

Let's now give some details about our version of DE strategy. The weight vectors evolve randomly with each pattern presentation (iteration) through the relation

$$v_i^{k+1} = w_i^k + \mu\left(w_{best}^k - w_i^k + w_{r1} - w_{r2}\right), \quad i = 1,\ldots NP, \quad (3)$$

where w_{best}^k is the best population member of the previous iteration, $\mu > 0$ is a real parameter (mutation constant) which regulates the contribution of the difference between weight vectors, and w_{r_1}, w_{r_2} are weight vectors randomly chosen from the population with $r_1, r_2 \in \{1,2,\ldots,i-1,i+1,\ldots,NP\}$, i.e. r_1, r_2 are random integers mutually different from the running index *i*. Aiming at increasing the diversity of the weight vectors further, a crossover-type operation is introduced that yields the so-called *trial* vector, $u_i^{k+1}, i = 1,\ldots NP$. This operation works as follows: the *mutant* weight vectors ($v_i^{k+1}, i = 1,\ldots NP$) are mixed with the *"target"* vectors, $w_i^{k+1}, i = 1,\ldots NP$. Specifically, we randomly choose a real

number r in the interval [0,1] for each component j, j=1,2,…, n, of the v_i^{k+1}. This number is compared with $\rho \in [0,1]$ (crossover constant), and if $r \leq \rho$ then the j-th component of the trial vector u_i^{k+1} gets the value of the j-th component of the mutant vector, v_i^{k+1}; otherwise, it gets the value of the j-th component of the target vector, w_i^{k+1}. The trial vector is accepted for the next iteration if and only if it reduces the value of the pattern-based error measure; otherwise the old value, w_i^k, is retained. This last operation is called *selection* and, due to the moving "optimum" nature of the on-line evolution task, it ensures that the fitness starts steadily decreasing at some iteration. The combined action of the mutation and crossover operation is responsible for much of the effectiveness of DE search, and allows DE strategies to act as parallel, noise-tolerant hill-climbing algorithms, which efficiently search the whole space for solutions [21].

EXPERIMENTS AND DISCUSSION

In our experiments, the colonoscopy video frames were separated into windows of size 16 by 16 pixels. Then the co-occurrence matrices algorithm was used to gather information regarding each pixel in an image window. Cooccurrence matrices represent the spatial distribution and the dependence of the grey levels within a local area. Based on these matrices, sets of statistical measures are computed for different angles. Four angles were considered, as well as a predefined distance of one pixel. The following four statistical measures provide high discrimination accuracy and were used to extract the feature vectors: *Energy-Angular Second Moment*, *Correlation*, *Inverse Difference Moment*, *Entropy*. The elements of these 16-dimensional feature vectors are the data that were presented to the network in order to train it (for a full description see [5]).

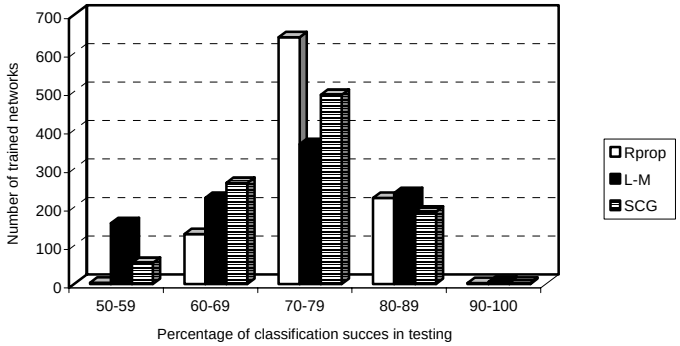

Figure 2: Generalisation results for three batch-training algorithms.

In the first set of experiments, one thousand 16-11-2 feed forward neural networks were trained off-line to detect malignant regions in one of the frames of a video sequence using the same training set of 300 patterns. The percentage of classification success in the test set (which included 3969 patterns from the same frame) for a set of 1000 trained networks is shown in Figure 2. One can observe that it is not easy to locate weights that will allow the networks to detect malignant regions with a success of over 90%. For example, in Figure 2, only 2 networks out of the 1000 trained with the Rprop algorithm, [22], achieved recognition success from 90% to 100%. For the Scaled Conjugate Gradient (SCG), [23], the corresponding number is 3 out of 1000, while for the Levenberg-Marquardt (L-M), [24], this number is slightly higher, as 6 out of the 1000 networks exhibited classification success between 90% and 100%. The best result for each training method is: 90% for the Rprop, 92.4% for the L-M and 92.6% for the SCG. However, a crucial factor for choosing a training algorithm is the average time of the training phase. In our case, average training times were: 0.644 secs for the Rprop; 13.056 secs for the L-M; 2.968 secs for the SCG.

In the second set of experiments, the Rprop algorithm (Rprop seems to be an attractive option for fast batch training) was compared to the classic on-line BP using data from another frame of the same video sequence. 300 patterns were used for training and 3969 for testing. The capability of the trained network (16-11-2 architectures were used) with the best performance in assigning appropriate characterisations (normal-cancer) to image regions is shown in Table 1.

Method	Cancer (%)	Normal (%)	Mean (%)
Rprop	83	96	93
On-line BP	73	93	88

Table 1: Best performance in terms of generalisation for Rprop and on-line BP.

The Rprop reveals, in general, a higher percentage of success than the on-line BP. The reader should, of course, keep in mind that Rprop minimises a batch error measure, i.e. it uses the true gradient of the error function as it exploits information from all the training patterns. The on-line BP, on the other hand, minimises a pattern-based error measure and works with an instantaneous approximation of the true gradient because information from only one pattern is used at each iteration. Therefore, on-line BP can be used for (re-)adapting to modified environment conditions, while Rprop requires all information about input-output patterns to be known a priori and, thus, fails to work when all the relevant features of the environment are not explicitly defined in advance. However, the results of the experiments presented so far make clear that the classical on-line BP needs further improvement in order to train networks for the accurate detection of malignant regions, or at least with comparatively success to batch training methods.

In the third experiment, a 16-11-2 architecture has been trained on-line to detect malignant regions in a set of four frames from the same video sequence. The frames used in the two previous experiments were included in the set. The network is initially trained on-line following the iterative scheme (1) for adapting the learning rate with the use of patterns from the first frame, and then differential evolution of learning occurs as data from the second frame appear at the input. The on-line evolution continuously adapts the network as patterns from other frames are presented in random order at the input. In total, 1200 patterns from the four frames of the video sequence were presented to the network during the training phase. The network was then tested using 15876 patterns from the four frames (4000 patterns approximately cover the whole image region of a frame and represent normal as well as malignant regions). The average capability of the trained networks in assigning appropriate characterisations to explored image regions is presented in Table 2.

Method	Frame 1	Frame 2	Frame 3	Frame 4
On-line BP	83%	84%	77%	88%
On-line BP seeded DE	93%	92%	84%	90%

Table 2: Average generalisation capability.

The on-line BP seeded DE scheme provides generalisation results close to the best results obtained by the batch training methods, as reported in the previous experiments. For example, the best SCG-trained network in the first experiment (trained off-line and tested using data from Frame 1) had 92.6% success, and the best Rprop-trained network in the second experiment (trained off-line and tested using data from Frame 2) had 93% success. With regards to the on-line BP, although the results of on-line BP in Table 2 are from networks that were trained and tested using data from only the corresponding frame, the performance of the method in terms of generalisation is not satisfactory. On the other hand, networks trained with on-line evolution are able to perform satisfactory in changing conditions, as data from different frames are presented to the same network. Thus, the on-line evolution trained network exhibits in all the cases better performance than the frame-specialised networks.

It is worth mentioning that the proposed training scheme seems to handle well the "catastrophic interference" among patterns of different frames, although further investigation is necessary to extract useful conclusions. Furthermore, in the reported experiments we gave no emphasis in fine-tuning the heuristic parameters of our scheme. Extensive testing with long video sequences and evaluation in interpreting more complex regions is necessary to fully investigate the properties, study the effect of the heuristic parameters and appreciate the performance of the hybrid learning strategy in this context.

ACKNOWLEDGEMENTS

The authors gratefully acknowledge the contribution of Dr. S. Karkanis, and Mr. D. Iakovidis of the *Digital Image Processing and Real-time Systems Lab* at the Department of Informatics and Telecommunications, University of Athens, Greece, in the acquisition of the video sequences and the extraction of the features.

REFERENCES

[1] Delaney, P.M; Papworth, G.D.; King, R.G., 1998, "Fibre optic confocal imaging (FOCI) for in vivo subsurface microscopy of the colon", in "Methods in disease: Investigating the gastrointestinal tract", Preedy V.R., Watson R.R. (eds.), Greenwich Medical Media, London, UK.

[2] Kwoh, C.K., 1995, "Probabilistic reasoning from correlated objective data", Ph.D. Thesis, Imperial College, London, UK.

[3] Looney, C.G., 1997, "Pattern recognition using neural networks", Oxford University Press, Oxford, UK.

[4] Innocent, P.R.; Barnes, M.; John, R., 1997, "Application of the fuzzy ART/MAP and MinMax/MAP neural network models to radiographic image classification", Artificial Intelligence in Medicine, 11, 241–263.

[5] Karkanis, S.; Magoulas, G.D.; Theofanous N., 2000, "Image recognition and neuronal networks: Intelligent systems for the improvement of imaging information", Minimal Invasive Therapy and Allied Technologies, 9, 225–230.

[6] Phee, S.J.; Ng, W.S.; Chen, I.M.; Seow-Choen, F.; Davies, B.L., 1998, "Automation of colonoscopy part II: Visual-control aspects", IEEE Engineering in Medicine and Biology, 81–88.

[7] Hanka, R.; Harte, T.P.; Dixon, A.K.; Lomas, D.J.; Britton, P.D., 1996, Neural networks in the interpretation of contrast-enhanced magnetic resonance images of the breast, Healthcare Computing 1996, Harrogate, UK, 275–283.

[8] Doulamis, A.D.; Doulamis, N.D.; Kollias, S.D., 2000, "On-line retrainable neural networks: improving the performance of neural networks in image analysis problems", IEEE Transactions on Neural Networks, 11, 137-155.

[9] Roehl, N.M.; Pedreira, C.E., 2001, "An online learning approach: a methodology for time varying applications", Neural Computing and Applications, 10, 101-107.

[10] Salomon, R.; Eggenberger, P., 1998, "Adaptation on the evolutionary time scale: a working hypothesis and basic experiments", 3rd European Conference on Artificial Evolution (AE'97), Nimes, France, Lecture Notes in Computer Science vol. 1363, Springer.

[11] Angeline, P., 1997, "Tracking extrema in dynamic environments", 6th Annual Conference on Evolutionary Programming VI, 335-345, Springer.

[12] Vavak, F.; Fogarty, T.C., 1996, "A comparative study of steady state and generational genetic algorithms", Evolutionary Computing: AISB Workshop, Lecture Notes in Computer Science vol. 1143, Springer.

[13] Plagianakos, V.P.; Magoulas, G.D.; Vrahatis, M.N., 2001, "Supervised training using global search methods", in "Advances in convex analysis and global optimisation", Hadjisavvas, N.; Pardalos, P. (eds.), vol. 54, Noncovex Optimization and its Applications, Kluwer Academic Publishers, Dordrecht, The Netherlands, pp.421-432.

[14] Sutton, R.S.; Whitehead, S.D., 1993, "Online learning with random representations", 10th International Conference on Machine Learning, Morgan Kaufmann, 314-321.

[15] Almeida, L.B.; Langlois, T.; Amaral, J.D.; Plankhov, A., 1998, "Parameter adaptation in stochastic optimisation", in "On-line Learning in Neural Networks", Saad, D. (ed.), 111-134, Cambridge University Press.

[16] Schraudolph, N.N., 1998, "Online local gain adaptation for multi-layer perceptrons", Technical Report, IDSIA-09-98, IDSIA, Lugano, Switzerland.

[17] Schraudolph, N.N., 1999, "Local gain adaptation in stochastic gradient descend", Technical Report, IDSIA-09-99, IDSIA, Lugano, Switzerland.

[18] Sutton, R.S., 1992, "Adapting bias by gradient descent: an incremental version of delta-bar-delta", 10th National Conference on Artificial Intelligence, MIT Press, 171-176.

[19] Saad, D., 1998, "On-line learning in neural networks", Cambridge University Press.

[20] Magoulas, G.D.; Plagianakos, V.P.; Vrahatis, M.N., 2001, "Adaptive stepsize algorithms for on-line training of neural networks", Nonlinear Analysis: Theory, Methods and Applications, 47, 3425-3430.

[21] Storn, R.; Price, K., 1997, "Differential evolution: a simple and efficient heuristic for global optimization over continuous spaces", Journal of Global Optimization, 11, 341-359.

[22] Riedmiller, M.; Braun, H., 1993, "A direct adaptive method for faster back-propagation learning: the Rprop algorithm", IEEE International Conference on Neural Networks, San Francisco, 586-591.

[23] Möller, M., 1993, "A scaled conjugate gradient algorithm for fast supervised learning", Neural Networks, 6, 525-533.

[24] Hagan, M.; Menhaj M., 1994, "Training feedforward networks with the Marquardt algorithm", IEEE Transcactions on Neural Nteworks, 5, 989-993.