ELSEVIER

Cancer Letters 96 (1995) 49–53

CANCER LETTERS

# Evolving neural networks for detecting breast cancer

David B. Fogel[a,*], Eugene C. Wasson III[b], Edward M. Boughton[c]

[a]*Natural Selection, Inc., 1591 Calle De Cinco, La Jolla, CA 92037, USA*
[b]*Maui Memorial Hospital, 221 Mahalani, Wailuku, HI 96793, USA*
[c]*Maui Economic Development Board, 590 Lipoa Pkwy., Kihei, Maui, HI 96753, USA*

Received 15 May 1995; accepted 18 May 1995

**Abstract**

Artificial neural networks are applied to the problem of detecting breast cancer from histologic data. Evolutionary programming is used to train the networks. This stochastic optimization method reduces the chance of becoming trapped in locally optimal weight sets. Preliminary results indicate that very parsimonious neural nets can outperform other methods reported in the literature on the same data. The results are statistically significant.

*Keywords:* Breast cancer; Neural networks; Evolutionary programming

## 1. Introduction

Artificial neural networks are parallel processing structures consisting of non-linear processing elements interconnected by fixed or variable weights. Such networks can be constructed to generate arbitrarily complex decision regions for stimulus-response pairs, thus they are well suited for use as pattern classifiers (readers unfamiliar with neural networks are referred to Ref. [1]). Although neural networks have been applied to difficult engineering problems with success, their application to problems in medicine has been limited. More specifically, with respect to diagnosis of breast cancer, only very recent research has been published in archived literature. These efforts have focused on extracting features from mammograms (e.g. [2,3]) or on results of laboratory blood tests and patient age (e.g. [4]), with a neural network being training on the features to indicate whether or not a malignancy is present.

The investigation of Wu et al. [2] used 43 preselected mammographic features related to abnormal density, microcalcification, parenchymal distortion, skin thickening, correlation with clinical findings, and so forth. Data was taken from 133 textbook cases in Ref. [5]. For each mammogram, each of the selected features was rated by an experienced mammographer on a scale of 0–10, and this served as the vector input to a multi-layer perceptron neural network. The network possessed 10 hidden units and a single output unit which was trained to yield a value of 0.0 for a benign case and 1.0 for a malignancy. Training was accomplished using back propagation. The results of this preliminary study and other described experiments indicated the suitability of this

[*] Corresponding author, Tel.: +1 619 4541590; Fax: +1 619 4541846; E-mail: fogel@sunshine.ucsd.edu.

0304-3835/95/$09.50
*SSDI* 0304-3835(95)03916-K

approach. By pruning the feature set to a more reasonable, smaller collection, the neural network was able to statistically significantly outperform attending radiologists and residents in assessing patterns of mammographic image features that are associated with benign and malignant lesions. There was no statistically significant difference between the performance of the network and the experienced mammographer used to rate each of the image features.

Floyd et al. [3] used back propagation on multilayer perceptrons to predict breast cancer from mammographic findings from patients who were scheduled for biopsy. They used only eight input parameters (mass size, mass margin, asymmetric density, architectural distortion, calcification number, calcification morphology, calcification density, and calcification distribution) and each of these was parameterized more objectively than in [2]. There were 260 cases used for training and testing. There was no held-out training set; all of the exemplars were processed in a jackknife statistical procedure. After significant training, the results indicated that if a threshold value of 0.1 were used, 38 out of 168 benign cases and all 92 malignancies would be identified. The authors compared this performance to that of radiologists and suggested that these results were statistically significantly better than radiologists at a $P < 0.08$ level. Although these results do appear fairly impressive with regard to detecting malignancy, the number of false positives appears rather high, and the statistical validity of the hypothesis test carried out can be questioned because the threshold of 0.1 was chosen after the authors reviewed the data and statistics were compiled on that same data. Thus the data did not reflect a random sample, but rather a biased sample. New data would have to be tested at the threshold value of 0.1 to have a sound statistical assessment.

Wilding et al. [4] used back propagation on multilayer perceptrons to assess both breast and ovarian cancer. Their procedure was similar to both [2] and [3] except that their input parameters consisted mainly of results of a group of laboratory blood tests (maximum of 10 total input parameters) from 104 patients. Unfortunately, Wilding et al. [4] report that the neural network was able to 'provide little improvement in the sensitivity of testing compared to the use of [the tumor marker] CA 15-3 only. Furthermore, it would appear that none of the networks appear to identify any worthwhile parameters or operating conditions with clinical utility.'

Each of these investigations used the back propagation training algorithm, which is essentially a gradient method that is well known to lead to suboptimal convergence at locally optimal weights sets. Thus, a network trained with back propagation may require many more hidden nodes to train to a tolerable level of error than are actually required. This excess in degrees of freedom subsequently hinders the generalization properties of the network, as it essentially overfits noise in the data. These concerns were specifically discussed in Refs. [3,4]. The most effective methods employed in these investigations for limiting the number of nodes and network parameters were based on sensitivity analysis and ad hoc pruning. Sensitivity analysis is problematic on non-linear transfer functions (such as neural networks) and ad hoc pruning can be largely unproductive. Despite directly mentioning concerns about overfitting their data, Floyd et al. [3] found that their best performance occurred when using 177 weights (16 hidden nodes), but they only had 260 samples. Wilding et al. [4] used networks with as few as 38 weights and as many as 132 weights, and despite having 104 samples were still unable to generate satisfactory performance. Even if the blood statistics that were being used were not particularly relevant to the classification task at hand, the failure to find suitable networks with more parameters than data may indicate the limitations of the training method.

## 2. Evolutionary programming

In contrast to back propagation, methods of evolutionary computation [6] can be used to train neural networks and evolutionary algorithms often outperform more classic optimization techniques. The procedures generally proceed as follows. A problem to be solved is cast in the form of an objective function that describes the worth of alternative solutions. Without loss of generality, suppose that the task is to find the solution that minimizes the objective function. A collection (population) of trial solutions are selected at random from some feasible range across the available parameters. Each solution is scored with respect to the objective function. The solutions (parents) are then mutated and/or recombined with other solutions in order to create new trials (offspring). These offspring are also scored with respect to the objective function and a subset of the parents and offspring are selected to become parents of the next iteration (generation) based on their relative performance. Those with superior performance are given a greater chance of being selected than are those of inferior quality. Fogel [6] details examples of evolutionary algorithms applied to a wide range of problems, including designing neural networks.

Designing neural networks through simulated evolution follows an iterative procedure:

1. A specific class of neural networks is selected. The number of input nodes corresponds to the amount of input data to be analyzed. The number of classes of concern (i.e. the number of output classification types of interest) determines the number of output nodes.

2. Exemplar data are selected for training.

3. A population of $P$ complete networks is selected at random. A network incorporates the number of hidden layers, the number of nodes in each of these layers, the weighted connections between all nodes in a feed-forward or other design, and all of the bias terms associated with each node.

4. Each of these 'parent' networks is evaluated on the exemplar data. A payoff function is used to assess the worth of each network. A typical objective function is the squared error between the target output and the actual output summed over all output nodes; this technique is often chosen because it simplifies calculations in the back propagation training algorithm. As evolutionary computation does not rely on similar calculations, any arbitrary payoff function can be incorporated into the process and can be made to reflect the operational worth of various correct and incorrect classifications.

5. 'Offspring' are created from these parent networks through random mutation. Simultaneous variation is applied to the values for the associated parameters (e.g. weights and biases of a multilayer perceptron). A probability distribution function is used to determine the likelihood of selecting combinations of these variations. The probability distribution can be preselected a priori by the operator or can be made to evolve along with the network, providing for nearly completely autonomous evolution [6].

6. The offspring networks are scored in a similar manner as their parents.

7. A probabilistic round-robin competition is conducted to determine the relative worth of each proposed network. Pairs of networks are selected at random. The network with superior performance is assigned a 'win'. Competitions are run to a preselected limit. Those networks with the most wins are selected to become parents for the next generation. In this manner, solutions that are far superior to their competitors have a corresponding high probability of being selected. The converse is also true. This function helps prevent stagnation at local optima by providing a parallel biased random walk.

8. The process iterates by returning to step 5.

The current investigation uses evolutionary programming to train multilayer feed-forward neural networks on data supplied to an internet anonymous ftp site (ics.uci.edu) by O. Mangasarian and collected by physician W.H. Wolberg, University of Wisconsin Hospitals. Previous publications with earlier versions of these data include [7,8]. The ftp site was last updated on July 15, 1992. These data consist of 699 instances of parameterized histopathology from breast biopsies. Each specimen has been assessed with regard to nine parameters: (1) clump thickness, (2) uniformity of cell size, (3) uniformity of cell shape, (4) marginal adhesion, (5) single epithelial cell size, (6) bare nuclei, (7) bland chromatin, (8) normal nucleoli, and (9) mitosis. Each parameter has been rated on an integer 10-scale. Of the 699 patterns, 458 are indicated as benign (65.5%) and 241 are indicated as malignant (34.5%).

Based on data supplied in January, 1989, consisting of 369 samples, Wolberg and Mangasarian [9] showed that when half of the data was used for training, 2 pairs of parallel hyperplanes applied to the parameterized data could achieve a level of 93.5% correct classification, and three pairs of such parallel hyperplanes could achieve a level of 95.9% correct. Zhang [10], operating on the same data, showed that a l-nearest neighbor technique could classify 93.7% of the data correctly (trained on 200 instances).

## 3. Method

Sixteen of the 699 data have missing values and were removed, leaving 683 data. The first 400 data were chosen as the training set while the remaining 283 were held out for testing. Two experimental designs were conducted. The first experiment consisted of five trials with a 9-2-1 multilayer perceptron (nine inputs, two hidden nodes, one output node) and five trials with a 9-9-1 multilayer perceptron. A population of 500 networks was evolved over 400 generations in each trial. Each parent was evaluated in terms of its total squared error (deviation from output to target value assigned to malignant or benign case). One offspring network was created from each parent at each generation. Each weight of each parent was mutated by adding a zero mean Gaussian random variable. The standard deviation of each mutation was set proportional to the parent's total squared error (i.e. the poorer the performance, the larger the mutation applied). After all of the offspring networks were scored, the networks underwent a tournament with 10 competitions (see above). The 500 networks with the most wins were selected to become parents of the next generation. The best network at the end of each trial was used to classify the test set. The second experiment extended the first to a sample size of 16 trials with a 9-2-1 multilayer perceptron based on results observed in the first experiment.

Table 1

Performance in the first experiment in which 500 networks were evolved over 400 generations using both nine and two hidden nodes

| Trial no. | MSE on training | Test Set errors and % correct |
|---|---|---|
| *9 Hidden nodes* | | |
| 1 | 0.12816 | 8 errors (6/2) 97.1% |
| 2 | 0.12484 | 7 errors (6/1) 97.5% |
| 3 | 0.12899 | 6 errors (5/1) 97.9% |
| 4 | 0.12605 | 7 errors (6/1) 97.5% |
| 5 | 0.12593 | 7 errors (5/2) 97.5% |
| *2 Hidden nodes* | | |
| 1 | 0.10409 | 6 errors (5/1) 97.9% |
| 2 | 0.11885 | 4 errors (3/1) 98.6% |
| 3 | 0.10389 | 4 errors (3/1) 98.6% |
| 4 | 0.10952 | 4 errors (3/1) 98.6% |
| 5 | 0.10397 | 5 errors (5/0) 98.2% |

The more parsimonious networks with only two hidden nodes generated superior training (400 samples) and test (283 samples) performance. The additional dimensions (100 versus 23) may have slowed training on the 9-9-1 networks. The $t$-test on the two-sample mean difference in testing is significant ($P < 0.002$) although the assumption of normality required for the test can be questioned. The numbers in parentheses indicate the number of false positives and false negatives, respectively. The percentages indicate the percent correct. MSE refers to mean squared error.

Table 2

Results of the second experiment comprising 16 independent trials with the 9-2-1 network

| | |
|---|---|
| Average MSE on training | 0.10841 (0.00461) |
| Average % correct on test set | 98.05 (0.46476) |
| Mean number of false positives | 4.25 (1.06458) |
| Mean number of false negatives | 1.25 (0.77460) |

The test set performance (283 samples) is statistically significantly superior to the best performance indicated in the literature with related data (95.9%) at $P < 0.05$ under non-parametric testing based on the Chebyshev inequality and $P << 0.001$ under a $t$-test. The numbers in parentheses indicate the sample standard deviations. MSE refers to mean squared error.

## 4. Results

The results of the first experiment are indicated in Table 1. After 400 generations in each of five trials, the average mean squared error on the training set was about 0.13 with the 9-9-1 network and about 0.11 with the 9-2-1. The test set performance appeared to favor the smaller network configuration (see Table 1 for discussion). The results of the second experiment are indicated in Table 2. The test set performance is statistically significantly superior to the best performance reported in the literature on related data from this archive.

## 5. Conclusions

These preliminary results indicate that a reasonably high level of performance can be achieved by small neural networks operating on histopathologic features from breast biopsies. The classification results presented on these data surpass the best results previously reported and routinely achieve greater than 97% correct. No explicit attempts were made to force the classification results to favor false positives, but the networks tended to err in that direction.

Although not considered here, the use of evolutionary programming for neural network training allows for implementing error functions that are not based on squared error. For example, strong arguments can be made that the cost of a false positive is not the same as a false negative. Evolutionary programming could be used to evolve classifier networks in the light of any payoff function, as opposed to back propagation which requires smooth error functions. Evolutionary optimization could also be used to design the topology of the network and select the input features. Investigation into these issues remains for future study.

The experimental design used in the current experiments does not rely on histologic data. The same procedure could be applied to parameterized mammograms, laboratory blood test results, or other suspected indicators of malignancy. Early detection of malignancy can greatly increase the chances for successful treatment and cure. Mammography is the current best method of screening for potential malignancies, but interpreting mammograms for the diagnosis of breast carcinoma is difficult because the radiologist must consider many related radiographic features of a suspicious lesion. There is currently considerable intra- and inter-observer disagreement or inconsistencies in mammographic interpretation.

The successful development of a neural network that is capable of reliably assessing the potential for the existence of breast carcinoma based on radiographic features of mammograms would make the radiologist both more efficient and more effective. To be truly effective, such an automated screening system would require the neural network to extract its own features from digitized mammograms, rather than rely on features extracted by the radiologist (cf. [2,3]). Initial attempts to train neural networks operating directly from film screen mammograms [11] have indicated modest success. Evolving such neural networks may lead to enhanced performance.

## Acknowledgments

Thanks to M.A. Williams and B. Barney for facilitating the use of the SP-2 at the Maui High Performance Computing Center. Research sponsored in part by the Phillips Laboratory, Air Force Materiel Command, USAF, under cooperative agreement number F29601-93-2-0001. The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Phillips Laboratory or the U.S. Government.

## References

[1] Haykin, S. (1994) Neural Networks: A Comprehensive Foundation. Macmillan, New York.

[2] Wu, Y.Z., Giger, M.L., Doi, K., Vybomy, C.J., Schmidt, R.A. and Metz, C.E. (1993) Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer. Radiology, 187, 81–87.

[3] Floyd, C.E., Lo, J.Y., Yun, A.J., Sullivan, D.C. and Kornguth, P.J. (1994) Prediction of breast cancer malignancy using an artificial neural network. Cancer, 74, 2944–2998.

[4] Wilding, P., Morgan, M.A., Grygotis, A.E., Shoffner, M.A. and Rosato E.F. (1994) Application of backpropagation neural networks to diagnosis of breast and ovarian cancer. Cancer Lett., 77, 145–153.

[5] Tabar, L. and Dean, P.B. (1985) Teaching Atlas of Mammography, 2nd edn. Thieme-Stratton, New York.

[6] Fogel, D.B. (1995) Evolutionary Computation: Toward a New Philosophy of Machine Intelligence. IEEE Press, Piscataway, NJ.

[7] Mangasarian, O.L. and Wolberg, W.H. (1990) Cancer diagnosis via linear programming. SIAM News, 23, 1–18.

[8] Bennett, K.P. and Mangasarian, O.L. (1992) Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods Software, 1, 22–34.

[9] Wolberg, W.H. and Mangasarian, O.L. (1990) Multisurface method of pattern separation for medical diagnosis applied to breast cytology. Proc. Natl. Acad. Sci. USA, 87, 9193–9196.

[10] Zhang, J. (1992) Selecting typical instances in instance-based learning. In: Proc. 9th Int. Machine Learning Conf., pp. 470–479. Editors: D. Sleeman and P. Edwards, Morgan Kaufmann, San Mateo, CA.

[11] Kocus, C.M., Rogers, S.K., Bauer, K.W., Steppe, J.M. and Hoffmeister, J.W. (1995) Neural network feature selection for breast cancer diagnosis. In: Applications and Science of Artificial Neural Networks, pp. 905–918. Editors: S.K. Rogers and D.W. Ruck. SPIE, Bellingham, WA.