An implementation of scatter search to train neural networks
for brain lesion recognition

Jeffrey Larson and Francis Newman

msp

# An implementation of scatter search to train neural networks for brain lesion recognition

Jeffrey Larson and Francis Newman

(Communicated by Kenneth S. Berenhaut)

In recent years, the use of computer aided diagnosis (CAD) has achieved acceptance in mammography and other areas. To facilitate automated detection of brain abnormalities, we propose a novel method for quickly training neural networks to classify brain images. Our method outperforms traditional neural network training methods by achieving a better balance between classification accuracy and training time.

## 1. Introduction

A variety of techniques have been implemented for lesion detection, including image filtering methods [Kotropoulos and Pitas 1992], support vector machines [Bilello et al. 2004], Markov random fields [Van Leemput et al. 2001], and a variety of artificial neural networks (ANNs) [Raff and Newman 1992; Wu et al. 1993; Yu and Guan 2000]. Nevertheless, automated pathology or lesion detection in most medical images has become somewhat dormant in recent years. Even in mammography, where computer-aided detection (CAD) has the greatest acceptance, the sensitivity (percentage of abnormal pathologies identified as "abnormal") is high, but the specificity (percentage of normal images where no abnormality is found) is poor. There are further difficulties for practical or commercial acceptance of CAD outside of mammography. These are often a combination of algorithmic and technical limitations. For example, when ANNs are used, backpropagation is frequently the method of choice for training. Backpropagation employs a multilayer feed forward architecture where error minimization is achieved via some form of gradient descent. Backpropagation (and its variants) can be slow to train, but a bigger problem for medical image diagnosis is the fact that the algorithm is likely to settle on an unsatisfactory local minimum [Gori and Tesi 1992; Sontag and Sussmann 1989].

Though the problem is difficult, correct classification of pathology and lesions in medical images could offer great benefits in reducing diagnostic errors and labor costs. It is estimated that 10-30% of breast cancers that are retrospectively visible are missed by radiologists upon initial reading [Brake et al. 1998], and 17-21% of polyps in computed tomography (CT) colonography are given false negative diagnoses due to human perception errors [Fletcher et al. 2000]. Missed tumors in lung CTs are disturbingly common [White et al. 1996], and brain tumors especially are frequently misdiagnosed [Wang et al. 2003]. Recent studies have found that physicians' clinical diagnoses are proven wrong 10-15% of the time by autopsy findings [Shojania et al. 2003; Roulson et al. 2005]. In an effort to reduce the number of misdiagnoses, researchers in the radiological sciences have been pursuing CAD since the early ascendancy of the computer [Schwartz 1970; Raff and Newman 1992; Chan et al. 1987]. Nevertheless, CAD has had limited impact in the field of human radiology, aside from some impact in mammography, where commercial systems have been available since 1998 [Vyborny et al. 2000].

In this effort we pursue a novel method to minimize the classification error for a feed-forward ANN in the medical image diagnosis problem using a scatter search meta-heuristic. The method, which is trained on a small subset of images and then validated on a larger set of images, greatly outperforms traditional classification methods.

## 2. Background

Scatter search [Glover 1999; Glover et al. 2000] is a population-based meta-heuristic that uses a local search algorithm to find an optimum. In this application, the population to be optimized is a set of random weights. Diverse individuals from the ANN population are then combined to form a "best" set according to some metric. This best set is then incorporated into the next population to be evaluated. The implementation of scatter search discussed here permits comparative evaluation of various feature vectors extracted from medical images.

For ease of reference, we define the terms to be used in this paper. An artificial neural network (ANN) is a mathematical model that simulates the structure of biological neural networks. It consists of interconnected nodes, each of which is a functional that acts on a linear combination of its inputs. Supervised ANNs are adaptive models that adjust the weights on the network arcs during a training phase in order to match each individual input to its target.

For example, a person might wish to construct a neural network to determine whether individuals should be diagnosed with colon cancer after their first screening. One might train the model using data (age, weight, height, ..., race, etc.) collected for 100 patients, as well as an indicator of whether or not they were diagnosed with
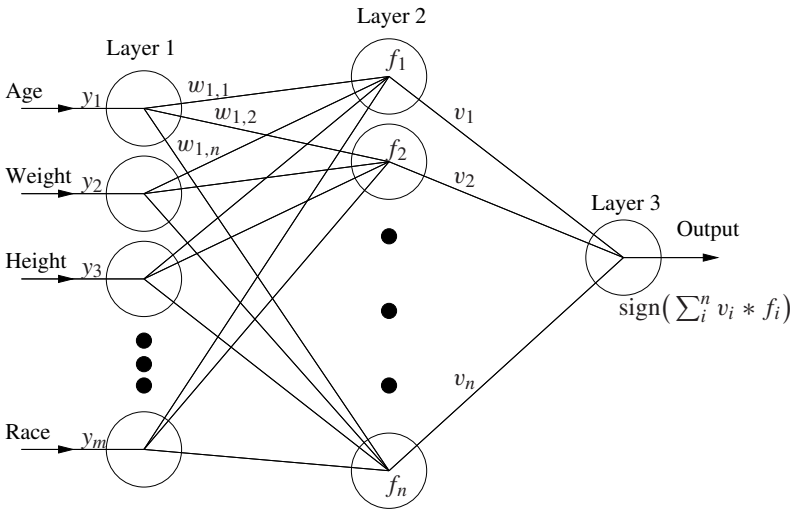
**Figure 1.** An example of an ANN.

colon cancer. To train the network, the weights on the arcs need to be adjusted until an acceptable percentage of the training set has the correct output (perhaps a network output of $+1$ for people who are diagnosed with cancer and $-1$ for those who are not). An example of such a three-layer network with $m$ inputs and $n$ hidden nodes is shown in Figure 1.

Our method employs a population of feed forward neural network architectures. According to the literature, three-layer networks are suitable for most problems, but the number of hidden nodes is frequently chosen by trial and error [Fausett 1994; Hassoun 1995, pp. 318–322]. It is important to note that our population consists of networks with identical three-layer architectures. Our process finds the network in the population that has the best weights and then uses these weights in an ANN to classify brain images. Though each network in our population has a fixed architecture with the same number of nodes, the method described in this paper permits easy and fast comparison of different hidden node architectures. Training a feed-forward ANN usually involves thousands of iterations (also known as epochs) to update the weights between layers in an effort to minimize the mean squared error. For a good treatment of neural networks, see [Fausett 1994; Hassoun 1995].

For a fixed architecture, training a neural network is a task in optimization. If a network has $n$ different weights, training involves finding $n$ weights that minimize the total error between the network output and the target values. Since backpropagation is a gradient descent method, its performance depends on initial network weights; if weights aren't well-initialized, backpropagation might perform poorly. Therefore, using a heuristic to find weights might yield a better neural network than backpropagation, in less time. Various heuristics have been proposed

to solve the problem of training neural networks [Kelly et al. 1996; Ye et al. 2007]. A heuristic such as scatter search, which explores many basins of attraction, could drastically outperform backpropagation.

Since this ANN is a supervised learning scheme, the a priori target values are available, and the algorithm seeks to minimize the error between the targets and the outputs for the training set. When the training returns an error below a particular threshold, the training is halted and a validation set is used for testing. (The threshold for the error term is problem and user-dependent.)

## 3. Implementation

**3.1.** *General scatter search implementation.* Scatter search is a population-based meta-heuristic, where a collection of preferred solutions are maintained and re-combined in order to generate new solutions. If the new solutions are preferred enough, they enter the population for the next iteration. For any given problem, the scatter search population has two subsets, good solutions and diverse solutions. The general framework for the algorithm is this:

(1) Generate a starting population.

(2) Perform a local search on every member of the population.

(3) Form a reference set of good solutions and diverse solutions using an appropriate metric.

(4) Form appropriate subsets from elements in the reference set.

(5) For each subset, generate new member(s) of the population.

(6) Return to step 2 and repeat until a satisfactory solution is found, or time runs out.

**3.2.** *Our implementation.* We attempt to find optimal weights for a neural network of fixed architecture; each instance is a fixed number of hidden nodes in a single layer. Each node in every implementation is a hyperbolic tangent activation function, which has been recommended as the best activation function for classification problems [Kalman and Kwasny 1992]. As with many meta-heuristics, any implementation allows for many degrees of freedom. For the purpose of classifying brain images as normal or abnormal, we adapted the scatter search algorithm as follows:

(1) Starting population: Generate 105 three layer networks with the same architecture. The weights (including biases) in each network are random numbers between $-1$ and 1.

(2) Local search: Perform a local search on the error function for each network using the Nelder–Mead method. From our initial weights, we are looking for a

local minimum of the total difference between the target output and the network output (for all training vectors).

(3) Form reference set: Select the 5 networks with the lowest total error to form the "good set". With this set fixed, we now create the "diverse set" by selecting networks that vary significantly from the good solutions. Select the network with the largest minimum Euclidean distance between its weights and the weights of the networks already in the reference set. Add this network to the reference set, and recalculate the minimum distance to the reference set for each network in the remaining population. Repeat this process until the reference set contains 10 networks: 5 good and 5 diverse.

(4) Form subsets of the reference set: For the 10 elements of the reference set, generate all possible unique pairs of networks. This creates $\binom{10}{2} = 45$ subsets.

(5) Generate new population: For each subset $\{x, y\}$, we generate three new elements of the population:

$$x_1 = x - v, \quad x_2 = x + v, \quad x_3 = y + v,$$

where $v = r(x - y)/2$ and $r$ is a random number between 0 and 1. Thus, for each $\{x, y\}$ pair, we are create 3 points on the line through $x$ and $y$. These 135 networks, along with 10 random networks to ensure diversity, form our new population. Also, the previous reference set is included in the population, though no local search needs to be performed on these networks.

## 4. Results

Normal and abnormal magnetic resonance images (MRIs) of the brain are used in this study (see Figure 2 for examples). Even with high-resolution images, representing individual images in a structure suitable for analysis is itself a considerable task. Though input of the entire image is desirable, the amount of data contained in a $256 \times 256$ gray-scale matrix is large. One widely used way to represent the data is to select regions of interest (ROIs) from an image; such an approach is widely used. Regions of interest are selected manually from each image and a single feature vector is generated for each region.

We examined a total of 250 normal images and 100 abnormal images. The neural network was trained on 10 of each type, while the remaining 330 were used as a validations set. To transform the images, we used the Haralick transform (a texture transform composed of second-order statistics) to generate feature vectors for each ROI. This transform uses the gray level co-occurrence matrix to uncover how often image pixel values appear adjacent to one another. If then computes quantities such as energy, correlation and homogeneity from that gray level co-occurrence matrix. For a complete discussion of the Haralick transform, see [Haralick et al. 1973].
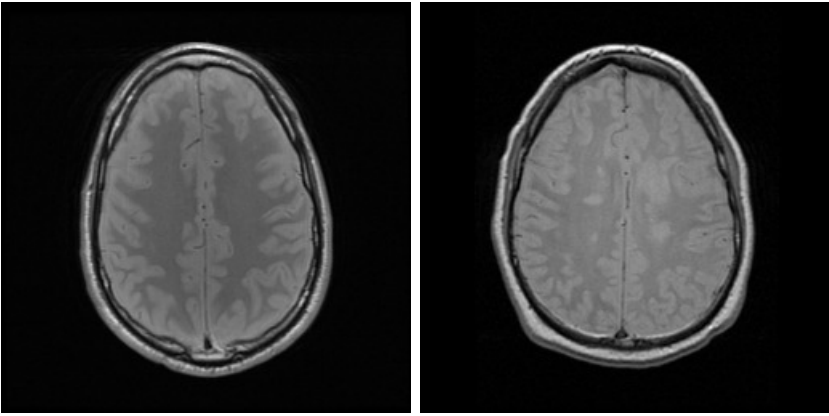
**Figure 2.** Normal (left) and abnormal (right) brain images.

Our training algorithm was used to classify 10 randomly selected normal images and 10 randomly selected abnormal images. Each network in the population consisted of 10 hidden nodes and 1 output node. With targets of $-1$ for normals and $+1$ for abnormals, a classification is considered successful when a given output is within 0.5 of the target. (Rounding is required since the output transfer function is a hyperbolic tangent function, and therefore $\pm 1$ is only reached at infinity.)

For comparison, we used the same training vectors to train an identical neural network architecture 10,000 times using back propagation (from 10,000 different random starting weights). With the various random starting points, different networks converged to a variety of weights. The "best" network (the one that best classified the training vectors) was validated using the remaining vectors. These results are summarized in Table 1.

| training set | training error | normals classified | abnormals classified | our classification rate | best back-propagation classification rate |
|---|---|---|---|---|---|
| random 10(a) | $8.23 \times 10^{-5}$ | 199 | 56 | 77.27% | 67.58% |
| random 10(b) | 2.48 | 185 | 48 | 70.61% | 59.39% |
| random 10(c) | 0.49 | 209 | 56 | 80.06% | 51.96% |
| random 10(d) | 4.32 | 206 | 53 | 78.48% | 55.76% |
| random 10(e) | 3.50 | 210 | 57 | 80.91% | 66.36% |
| random 10(f) | 0.93 | 178 | 83 | 79.09% | 62.12% |

**Table 1.** Comparison of classification results between our training method and the best network (out of 10,000) trained using backpropagation for 6 different randomly selected training sets.

## 5. Discussion

The results reflect a dramatic improvement in the classification rate compared to that of backpropagation. This improvement is surprising given that the local search for our implementation is rudimentary. One would expect a rigorous local search, similar to what is used in backpropagation, to outperform Nelder–Mead. While there may not be a logical explanation for why taking a linear combination of the weights from two decent networks results in a worthwhile network, generating our population in the manner prescribed certainly allows us to identify many different basins of attraction. The fact that these results were achieved with only 20 training vectors is even more surprising since the number of training vectors usually used for the medical image recognition problem is orders of magnitude greater than 20 [Baum and Haussler 1989]. Expanding the training set would likely improve the classification rate.

The proposed method has many strengths. For example, the error function can easily be varied for different applications. In this implementation, false positives and false negatives were weighted equally. If sensitivity is more important than specificity, for example, it might be preferential to weight the error corresponding to missed abnormals higher than misclassified normals. Also, this training method has an ease of implementability. Although different network architectures (number of hidden nodes, layers, activation functions, etc.) might be better suited for different problem classes, this algorithm allows for quick testing of different networks. Changing any network parameter is simple, and training these different networks can be accomplished in a few minutes. Interested parties can receive the MATLAB code used in our implementation by emailing the corresponding author.

This success opens a number of avenues for further exploration. For example, the number of hidden nodes can be varied to determine whether a larger or smaller network better suits a given problem. A desire for the ability to meaningfully compare different feature vectors has been expressed in the literature [Duda et al. 2001; Egmont-Petersen et al. 2002], and the proposed network training algorithm can facilitate this well. Aside from feature vectors, different sizes of networks, different activation functions, and different network architectures can all easily be tested and compared with this algorithm.

## 6. Conclusion

In this paper, we propose a novel method for training neural networks for the specific task of classifying medical images as normal or abnormal. Our proposed method shows great promise for this task, but also has an ease of implementation that allows for quick training of neural networks for general classification problem.

## Acknowledgments

## References

[Baum and Haussler 1989] E. B. Baum and D. Haussler, "What size net gives valid generalization?", *Neural Computation* **1**:1 (1989), 151–160.

[Bilello et al. 2004] M. Bilello, S. B. Gokturk, T. Desser, S. Napel, J. R. Brooke Jeffrey, and C. F. Beaulieu, "Automatic detection and classification of hypodense hepatic lesions on contrast-enhanced venous-phase CT", *Medical Physics* **31**:9 (2004), 2584–2593.

[Brake et al. 1998] G. Brake, N. Karssemeijer, and J. Hendriks, "Automated detection of breast carcinomas not detected in a screening program", *Radiology* **207**:2 (1998), 465–471.

[Chan et al. 1987] H.-P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon, and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. I. Automated detection of microcalcifications in mammography", *Medical Physics* **14**:4 (1987), 538–548.

[Duda et al. 2001] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification*, 2nd ed., Wiley-Interscience, New York, 2001. MR 2001k:68120 Zbl 0968.68140

[Egmont-Petersen et al. 2002] M. Egmont-Petersen, D. de Ridder, and H. Handels, "Image processing with neural networks: a review", *Pattern Recognition* **35**:10 (2002), 2279–2301.

[Fausett 1994] L. Fausett, *Fundamentals of neural networks: architectures, algorithms, and applications*, Prentice Hall, 1994. Zbl 0828.68107

[Fletcher et al. 2000] J. G. Fletcher, C. D. Johnson, T. J. Welch, R. L. MacCarty, D. A. Ahlquist, J. E. Reed, W. S. Harmsen, and L. A. Wilson, "Optimization of CT colonography technique: prospective trial in 180 patients", *Radiology* **216**:3 (2000), 704–711.

[Glover 1999] F. Glover, "Scatter search and path relinking", pp. 297–316 in *New ideas in optimization*, edited by D. Corne et al., McGraw-Hill, Maidenhead, UK, 1999.

[Glover et al. 2000] F. Glover, M. Laguna, and R. Martí, "Fundamentals of scatter search and path relinking: Tabu search for combinatorial optimization", *Control Cybernet.* **29**:3 (2000), 653–684. MR 1835135 Zbl 0983.90077

[Gori and Tesi 1992] M. Gori and A. Tesi, "On the problem of local minima in backpropagation", *IEEE Trans. Pattern Anal. Machine Intel.* **14** (1992), 76–86.

[Haralick et al. 1973] R. M. Haralick, K. Shanmugam, and I. Dinstein, "Textural features for image classification", *IEEE Trans. Systems Man Cybernetics* **3**:6 (1973), 610–621.

[Hassoun 1995] M. H. Hassoun, *Fundamentals of artificial neural networks*, The MIT Press, Cambridge, MA, 1995. Zbl 0850.68271

[Kalman and Kwasny 1992] B. L. Kalman and S. C. Kwasny, "Why tanh: choosing a sigmoidal function", pp. 578–581 in *International Joint Conference on Neural Networks*, vol. 4, 1992.

[Kelly et al. 1996] J. Kelly, B. Rangaswamy, and J. Xu, "A scatter search-based learning algorithm for neural network training", *J. Heuristics* **2** (1996), 129–146. Zbl 0873.68182

[Kotropoulos and Pitas 1992] C. Kotropoulos and I. Pitas, "Optimum nonlinear signal detection and estimation in the presence of ultrasonic speckle", *Ultrasonic Imaging* **14**:3 (1992), 249–275.

[Raff and Newman 1992] U. Raff and F. D. Newman, "Automated lesion detection and lesion quantitation in MR images using autoassociative memory", *Medical Physics* **19**:1 (1992), 71–77.

[Roulson et al. 2005] J. Roulson, E. W. Benbow, and P. S. Hasleton, "Discrepancies between clinical and autopsy diagnosis and the value of post mortem histology; a meta-analysis and review.", *Histopathology* **47**:6 (2005), 551–559.

[Schwartz 1970] W. B. Schwartz, "Medicine and the computer: The promise and problems of change", *New England J. Medicine* **283**:23 (1970), 1257–1264.

[Shojania et al. 2003] K. G. Shojania, E. C. Burton, K. M. McDonald, and L. Goldman, "Changes in rates of autopsy-detected diagnostic errors over time: a systematic review", *J. Amer. Medical Assoc.* **289**:21 (2003), 2849–2856.

[Sontag and Sussmann 1989] E. D. Sontag and H. J. Sussmann, "Backpropagation can give rise to spurious local minima even for networks without hidden layers", *Complex Systems* **3**:1 (1989), 91–106. MR 92a:68127 Zbl 0732.68091

[Van Leemput et al. 2001] K. Van Leemput, F. Maes, D. Vandermeulen, A. Colchester, and P. Suetens, "Automated segmentation of multiple sclerosis lesions by model outlier detection", *IEEE Trans. Medical Imaging* **20**:8 (2001), 677–688.

[Vyborny et al. 2000] C. J. Vyborny, T. Doi, K. F. O'Shaughnessy, H. M. Romsdahl, A. C. Schneider, and A. A. Stein, "Breast cancer: importance of spiculation in computer-aided detection", *Radiology* **215**:3 (2000), 703–707.

[Wang et al. 2003] J. Wang, T. Bo, I. Jonassen, O. Myklebost, and E. Hovig, "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data", *BMC Bioinformatics* **4**:1 (2003), 60.

[White et al. 1996] C. S. White, B. M. Romney, A. C. Mason, J. H. Austin, B. H. Miller, and Z. Protopapas, "Primary carcinoma of the lung overlooked at CT: analysis of findings in 14 patients", *Radiology* **199**:1 (1996), 109–115.

[Wu et al. 1993] Y. Wu, M. L. Giger, K. Doi, C. J. Vyborny, R. A. Schmidt, and C. E. Metz, "Artificial neural networks in mammography: application to decision making in the diagnosis of breast cancer", *Radiology* **187**:1 (1993), 81–87.

[Ye et al. 2007] J. Ye, J. Qiao, M.-a. Li, and X. Ruan, "A tabu based neural network learning algorithm", *Neurocomput.* **70**:4-6 (2007), 875–882.

[Yu and Guan 2000] S. Yu and L. Guan, "A CAD system for the automatic detection of clustered microcalcifications in digitized mammogram films", *IEEE Trans. Medical Imaging* **19**:2 (2000), 115–126.

jeffrey.larson@ucdenver.edu    *University of Colorado Denver, Department of Mathematical and Statistical Sciences, Campus Box 170, P.O. Box 173364, Denver, CO 80217-3364, United States*

francis.newman@ucdenver.edu    *University of Colorado Denver, Department of Radiation Oncology, 1665 Aurora Court, Box F-706, Aurora, CO 80045, United States*

# involve

msp.berkeley.edu/involve

See inside back cover or http://msp.berkeley.edu/involve for submission instructions.

The subscription price for 2011 is US $100/year for the electronic version, and $130/year (+$35 shipping outside the US) for print and electronic. Subscriptions, requests for back issues from the last three years and changes of subscribers address should be sent to Mathematical Sciences Publishers, Department of Mathematics, University of California, Berkeley, CA 94704-3840, USA.

# involve

2011     vol. 4     no. 3