CENTER FOR
PARALLEL OPTIMIZATION

NEURAL NETWORK TRAINING VIA LINEAR PROGRAMMING

by

Kristin P. Bennett & Olvi L. Mangasarian

Computer Sciences Technical Report #948

July 1990

Neural Network Training via Linear Programming

Kristin P. Bennett †

Olvi L. Mangasarian †

Abstract. An efficient algorithm for training a feed-forward neural network with partially pre-assigned weights is proposed. The algorithm is based on linear programming and has a number of advantages over back propagation such as: automatic determination of the number of hidden units, 100% correctness on the training set, faster training, and elimination of parameters from the algorithm. The proposed method is currently in use for breast cancer diagnosis.

1. Introduction

We propose a fast polynomial-time approach for training neural networks based on linear programming. The principal idea is based on the Multisurface Method (MSM) for the separation of two disjoint pattern sets with intersecting convex hulls [Mang68]. The Multisurface Method which is currently in use for breast cancer diagnosis [Mang89, Wolb90], can be modeled as a feed-forward neural network with a partially-fixed topology. Training neural networks by MSM has the following advantages over the Back Propagation Algorithm (BP) [Rume86]:

(a) Automatic determination of the number of hidden units.

(b) Achieving 100% correctness on the training set.

(c) Faster training.

(d) Elimination of parameters from the training algorithm.

We begin with a description of the MSM linear programming approach for training a neural network for the binary classification of two disjoint point sets. Extension of MSM to discrimination between more than two sets can be achieved by $\log_2 s$ binary classifications where s is the number of sets. For simplicity of presentation we confine ourselves to classifying two sets only.

2. The MSM Classifier

Let the the finite pattern sets **A** and **B** be two given disjoint training sets in the n-dimension real feature space R^n. Let the cardinality of **A** and **B** be m and k respectively. The sets **A** and **B** are represented by the $m \times n$ and $k \times n$ matrices A and B. If the convex hulls of the sets **A** and **B** do not intersect, or equivalently if they are linearly

† Computer Sciences Department, University of Wisconsin, 1210 West Dayton Street, Madison, WI 53706. Research supported by National Science Foundation Grants DCR-8521228 and CCR-8723091, Air Force Office of Scientific Research Grant AFSOR 89-0410, and Air Force Laboratory Graduate Fellowship S-789-000-053.

separable, a single linear program can generate a separating plane in polynomial time [Karm84, Khac79] by solving the following problem [Mang65]:

$$\underset{\alpha,\beta,w}{\text{maximize}} \left\{ \alpha-\beta \mid Aw \geq e\alpha,\ Bw \leq e\beta,\ -e \leq w \leq e \right\} \tag{1}$$

where $w \in R^n$ is the weight vector associated with the separating plane, $\frac{\alpha+\beta}{2}$ is the threshold that locates the separating plane, and e is a vector of ones in a real space of arbitrary finite dimension. Note that this linear program generates the weight vector w and threshold $\frac{\alpha+\beta}{2}$ for a linear threshold unit which discriminates between two linearly separable sets.

When the sets **A** and **B** are not linearly separable, $w=0, \alpha=\beta=0$ is a solution to (1), and no useful information is derived from the problem. Hence problem (1) must be modified to generate a sequence of different pairs of separating planes which constitute the MSM classifier. Each pair of parallel planes distinguishes a subset of **A** from a subset of **B**. Total separation is achieved by the piecewise-linear MSM classifier which we now describe.

The piecewise-linear surface consists of p pairs of planes with weight vectors $w^1, \ldots, w^p \in R^n$, and thresholds $-\alpha^1, \ldots, -\alpha^{p-1}, \frac{\alpha^p+\beta^p}{2}$, and $\beta^1, \ldots, \beta^{p-1}, \frac{\alpha^p+\beta^p}{2}$, with $\alpha^i<\beta^i, i=1,\ldots,p-1$ and $\alpha^p>\beta^p$. Classification is achieved as follows:

MSM Classifier for $x \in R^n$

for i = 1 to p-1
begin
  if $xw^i>\beta^i$ then $x \in A$; stop
  if $-xw^i>-\alpha^i$ then $x \in B$; stop
end
if $xw^p \geq \frac{\alpha^p+\beta^p}{2}$ then $x \in A$
else $x \in B$

Geometrically the MSM classifier corresponds to separation by a piecewise-linear surface as illustrated in Figure 1 for a hypothetical case of two pairs of planes in two dimensions. Figure 2 depicts separation of the Wisconsin Breast Cancer Data (WBCD), the actual clinical data described in Section 5 of this paper. This figure shows which points are separated by a given pair of planes and which points remain to be separated by a succeeding pair of planes. The top graph in Figure 2 is the projection of 369 nine-dimensional data points on the two-dimensional space spanned by the normals, w^1 and w^2, to the first two pairs of separating planes. Note that w^1 and w^2 are not orthogonal to each other in general. In the bottom graph, the points which were not classified by the first

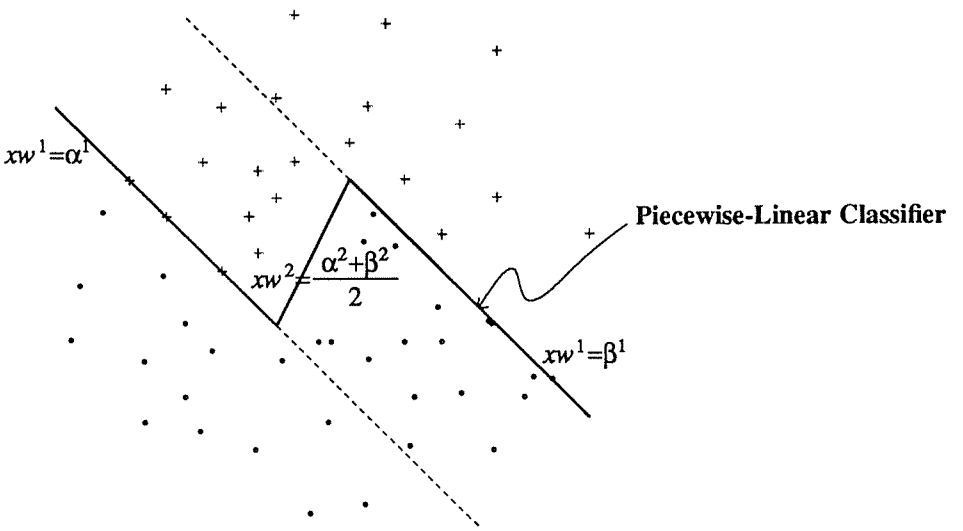


Fig 1: Geometric Depiction of MSM Classifier

two pairs of parallel planes (i.e. the points in the parallelogram formed in the top graph) are projected on the two-dimensional space spanned by the normals, w^3 and w^4, to the last two pairs of separating planes. Complete separation of the points is achieved in the bottom graph.

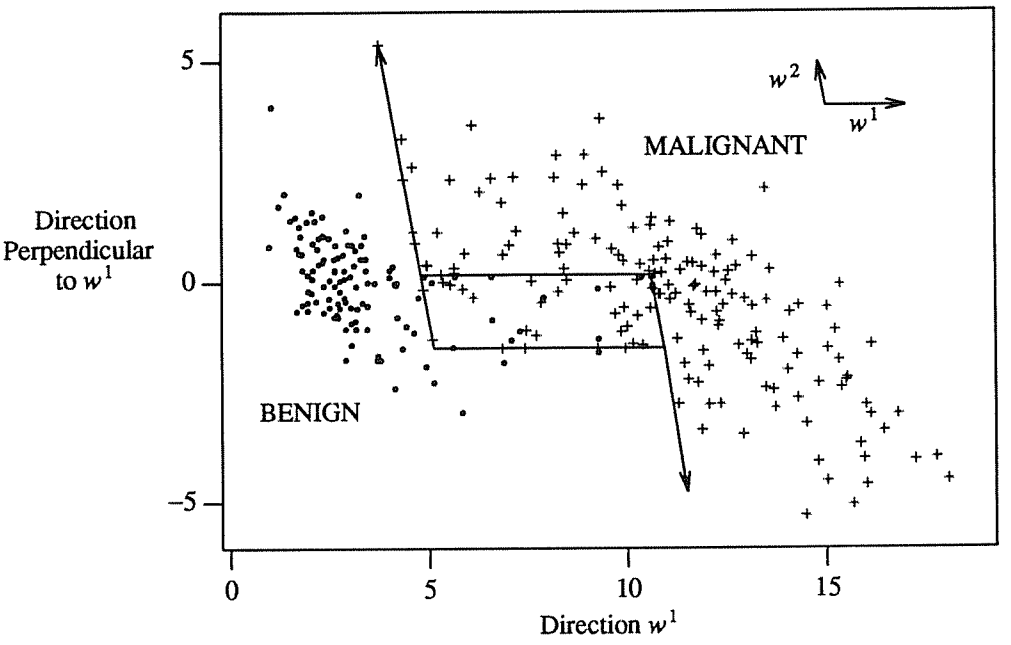
3. MSM Classifier as a Neural Network

We give now a novel representation of the MSM classifier as a trained feed-forward neural network which is depicted in Figure 3 and which can be trained efficiently by our linear programming approach. This network is composed of n input units, $2p-1$ hidden units, and 1 output unit. For $i=1,...,p-1$, w^i and $-w^i$ are the incoming weights to the $(2i-1)^{th}$ and the $(2i)^{th}$ hidden units with thresholds β^i and $-\alpha^i$ respectively. For $i=p-1$, w^p is the incoming weight to the $(2p-1)^{th}$ hidden unit with threshold $\frac{\alpha^p+\beta^p}{2}$. The weights on the arcs connecting the $2p-1$ hidden units to the output unit (see Figure 3) are predetermined such that the activation of the output unit is caused by the firing hidden unit with the lowest index. The threshold of the output unit is 0. To reduce clutter in Figure 2, the n arcs connecting the n input units to a hidden unit are consolidated into one arc.

Note that predetermined weights are used between the hidden units and the output unit. This may restrict the representational power of the network for a fixed number of units. However this does not seem to degrade the performance of the MSM neural network. MSM dynamically determines the number of hidden units required in order to correctly classify all the training examples. The performance of MSM is comparable to unrestricted networks trained with BP. In fact, the number of hidden units determined by MSM is a good estimate of the number required

WBCD Points separated by 1^{st} and 2^{nd} pairs of planes

Space Spanned by Directions w^1 and w^2



WBCD Points Separated by 3^{rd} and $4th$ Pairs of Planes

Space Spanned by Direction w^3 and w^4

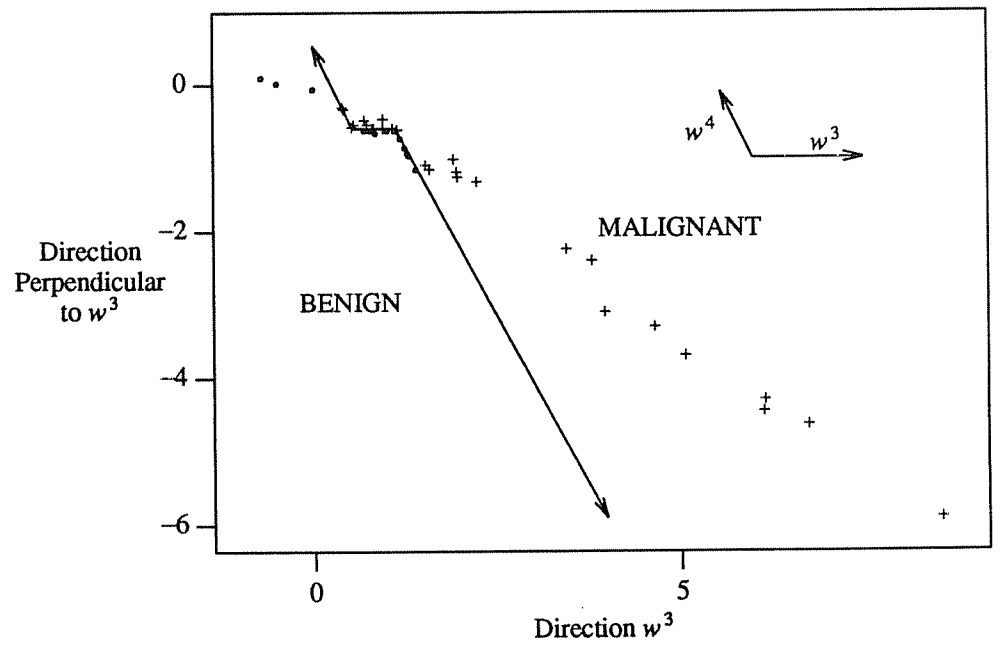


Fig 2: MSM Classifier for Wisconsin Breast Cancer Data

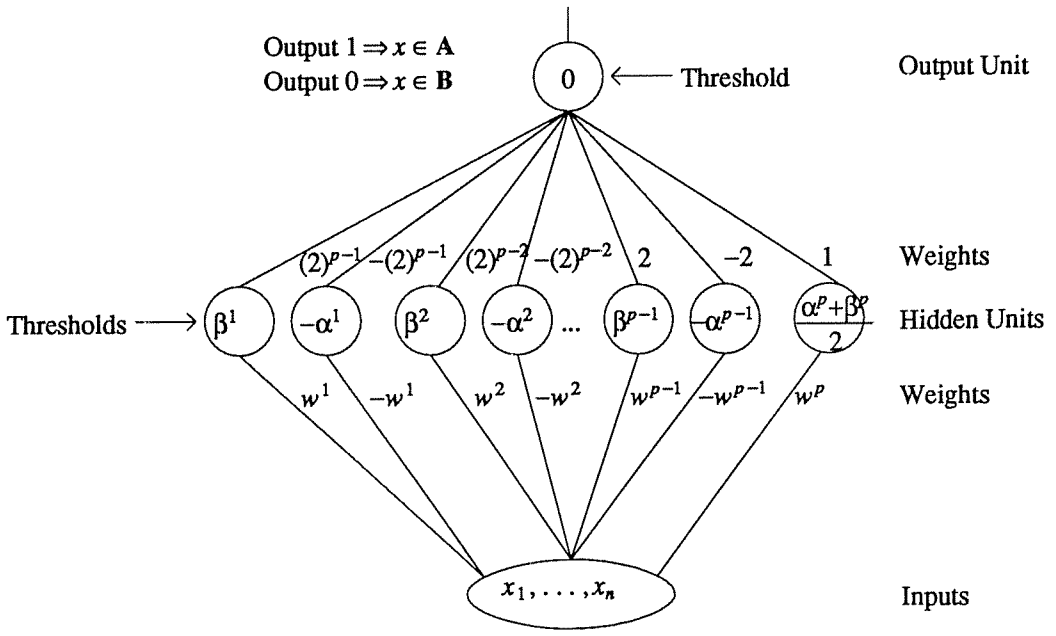


Fig 3. MSM Classifier as a Neural Network

for a given classification task by an unconstrained network using BP. For example, consider the n-bit parity function which takes an n-bit binary vector as its argument and returns a 1 if the number of one's is odd, and 0 otherwise. This problem requires n hidden units for a feed-forward network [Rume86]. In our tests the MSM topology required n hidden units when n is odd, and n+1 units when n is even, and the network was trained in considerably less time using MSM than with BP. Figure 4 plots the execution time of MSM and BP on several n-bit parity problems on a DECstation 3100.

We note that generating a trained feed-forward network to distinguish between the disjoint pattern sets **A** and **B** can be posed as the following nonlinear nonconvex optimization problem which is solved exactly by MSM:

For some integer p:

$$\underset{\substack{W\in R^{n\times p},\\ \alpha\in R^p,\ \beta\in R^p\\ \beta_i>\alpha_i,\ i=1,\ldots,p-1,\\ \beta_p=\alpha_p}}{\text{minimum}} \left[\left\|\left(\sum_{i=1}^{p} -2^{p-i}\left((AW_{\cdot i}-e\beta_i)_*-(-AW_{\cdot i}+e\alpha_i)_*\right)+\tfrac{1}{2}e\right)_*\right\|_1+\left\|\left(\sum_{i=1}^{p} 2^{p-i}\left((BW_{\cdot i}-e\beta_i)_*-(-BW_{\cdot i}+e\alpha_i)_*\right)+\tfrac{1}{2}e\right)_*\right\|_1\right]=0 \quad (2)$$

where $((d)_*)_j := \begin{cases}1 & \text{if } d_j>0\\ 0 & \text{if } d_j\le 0\end{cases}$, e is a vector of ones, $\|\cdot\|_1$ denotes the 1-norm,
$R^{n\times p}$ is the space of $n\times p$ real matrices, and $W_{\cdot i}$ is the i^{th} column of the matrix W.

The above function counts the number of misclassified points. The summation inside the first norm counts the number of points in **A** that are misclassified. The summation within the second norm counts the number of the misclassified points of **B**. To understand how the counting works, we examine the relationship between the above

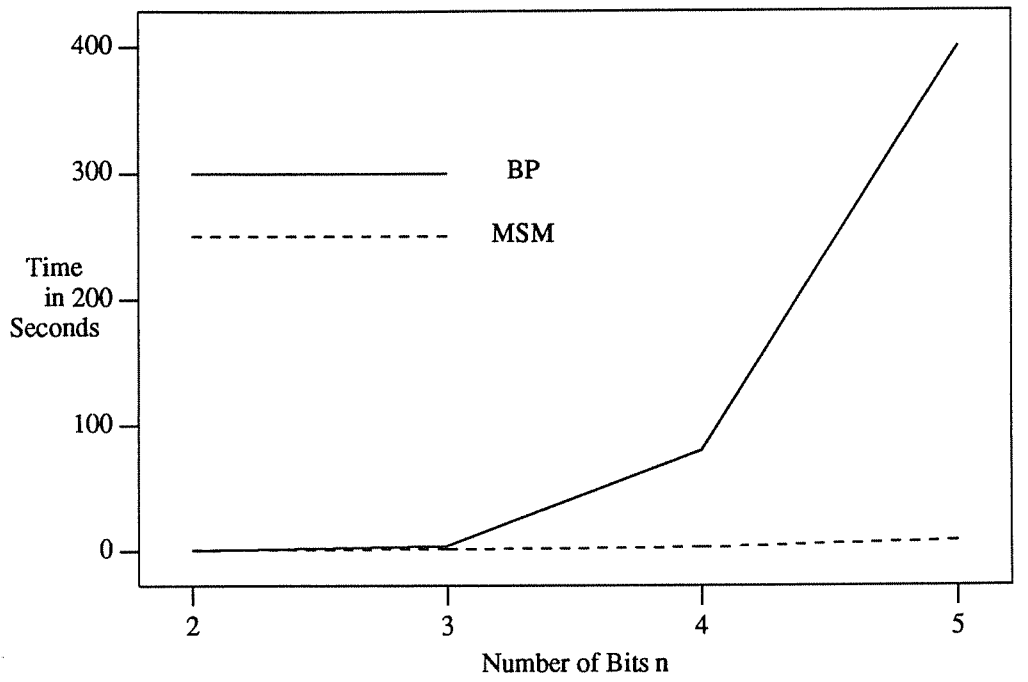


Fig 4: Execution Times for BP and MSM on n-Bit Parity Problems

function and the MSM classifier. The vector $W_{.i}$ is the normal to the i^{th} pair of separating planes within the MSM classifier, and α_i, and β_i locate these planes. Consider the j^{th} point in **A** represented by the j^{th} row A_j of the matrix A. Let A_j be correctly classified by the l^{th} pair of planes, then the following is true:

- A_j lies on or between the pairs of parallel planes $xW_{.k}=\alpha_k$ and $xW_{.k}=\beta_k$ for $k=1,...,l-1$ which implies that $\alpha_k \le A_j W_{.k} \le \beta_k$ for $k=1,...,l-1$. Thus:

$$\sum_{i=1}^{l-1} -2^{p-i}\left[(A_j W_{.i}-\beta_i)_*-(-A_j W_{.i}+\alpha_i)_*\right] = 0 \tag{3}$$

- A_j is in the open halfspace $\left\{x \mid xW_{.l}>\beta_l>\alpha_l\right\}$. Thus:

$$-2^{p-l}\left[(A_j W_{.l}-\beta_l)_*-(-A_j W_{.l}+\alpha_l)_*\right] = -2^{p-l} \tag{4}$$

- A_j maybe "anywhere" in relation to the remaining pairs of parallel planes: $xW_{.i}=\alpha_i$ and $xW_{.i}=\beta_i$, $i=l+1,...,p$. The coefficients attached to the remaining planes in (2) are chosen such that the combined effect of these planes will not cause A_j to be counted as misclassified. More specifically:

$$\sum_{i=l+1}^{p} -2^{p-i}\left[(A_j W_{.i}-\beta_i)_*-(-A_j W_{.i}+\alpha_i)_*\right] \le \sum_{i=l+1}^{p} -2^{p-i}(0-1) = 2^{p-l}-1 \tag{5}$$

Summing the above equalities implies that

$$\sum_{i=1}^{p} -2^{p-i}\left[(AW_{.i}-\beta_i)_*-(-AW_{.i}+\alpha_i)_*\right] + \tfrac{1}{2} \le -\tfrac{1}{2} \tag{6}$$

and hence:

$$\left[\sum_{i=1}^{p} -2^{p-i}\left[(AW_{.i}-\beta_i)_*-(-AW_{.i}+\alpha_i)_*\right] + \tfrac{1}{2}\right]_* = 0 \tag{7}$$

Thus the first part of the function in (2) correctly returns 0 for A_j.

If A_j is misclassified at the l^{th} plane, the equation (3) above remains the same. The right-hand side of (4) becomes 2^{p-l}. Equation (5) becomes:

$$\sum_{i=l+1}^{p} -2^{p-i}\left[(A_jW_{.i}-\beta_i)_*-(-A_jW_{.i}+\alpha_i)_*\right] \ge \sum_{i=l+1}^{p} -2^{p-i}(1-0) = -2^{p-l}+1 \tag{5'}$$

Adding equations (3), the modified equation (4), and (5'), we get:

$$\sum_{i=1}^{p} -2^{p-i}\left[(AW_{.i}-\beta_i)_*-(-AW_{.i}+\alpha_i)_*\right] + \tfrac{1}{2} \ge \frac{3}{2} \tag{6'}$$

Hence when $A_{.j}$ is misclassified, the expression in (7) equals 1, and the first part of (2) returns 1 for A_j. Therefore the first norm of (2) counts the number of misclassified points of the set **A**. Similarly the second norm in (2) counts the number of misclassified points of the set **B**. The term $\frac{1}{2}e$ in (2) is added in order to eliminate the ambiguous case of points of either set **A** or **B** lying on the final separating plane $xW_{.p}=\alpha_p=\beta_p$. Note that no points are misclassified if and only if the right side of (2) equals 0. In addition, the solution of (2) is an MSM classifier, and a solution for (2) such that all points are correctly classified can always be found using MSM.

4. Training the MSM Neural Network

In order to generate the MSM classifier for the general case of linearly inseparable pattern sets, problem (1) is modified as follows to ensure that w is nonzero:

$$\underset{\alpha,\beta,w}{\text{maximize}}\left\{\alpha-\beta \mid Aw\ge e\alpha,\ Bw\le e\beta,\ -e\le w\le e,\ w\ne 0\right\} \tag{8}$$

Each problem (8) is solved by a finite sequence of linear programs. This sequence does not exceed 2n and is often less depending on the way that the constraint $w\neq 0$ is implemented [Mang68, Mang89]. Each problem (8) separates portions of the sets **A** and **B** from each other. The points thus separated are removed from **A** and **B**. This process is repeated until no points remain. The algorithm with an antidegeneracy procedure ensures that total separation of any two disjoint point sets can be achieved in polynomial time.

Note that unlike BP, there are no parameters in the learning algorithm that must be determined experimentally in MSM. For example, the number of hidden units within the MSM neural net is determined automatically by the program.

5. Computational Comparison of BP and MSM on Medical Diagnosis Problems

MSM is currently in active use at University of Wisconsin Hospitals for the diagnosis of breast cancer [Wolb90, Mang89]. We briefly discuss this application first and give comparisons with BP.

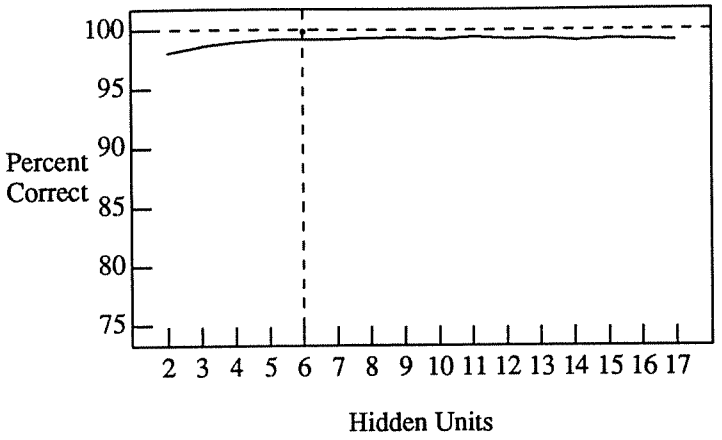
The Wisconsin Breast Cancer Data (WBCD) set, developed by Dr. W. H. Wolberg, consists of nine measurements taken from fine needle aspirates from human breast tissue. On this data set, MSM was trained originally on 369 samples and was tested subsequently on 70 newly acquired samples all of which were classified correctly except one. At that time it was retrained and since then it has correctly classified all 48 subsequent samples. For reliability in medical diagnosis applications, 100% correctness on the training and testing sets is very important. Such levels of training set correctness were attained by MSM on this data set but not by other approaches such as BP, statistical pattern separation and decision tree approaches [Wolb88]. It is important to emphasize that in medical diagnosis, classification is performed on all available data and the classifier is then used on incoming data. Table 1 compares the best results obtained by BP as implemented in [McCl87] with the results from MSM on the WBCD. It is interesting to note that BP does not achieve the correctness rate on either training or testing sets achieved by MSM.

	MSM	BP
Training Time (seconds)	108	469.5
Training Set Correctness (%)	100.0	98.9
Testing Set Correctness (%)	98.3	94.9
Number of Hidden Units	7	6

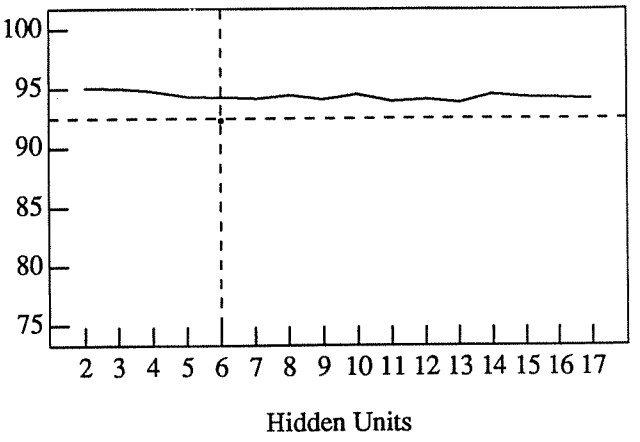
Table 1: Comparison of MSM and BP on Wisconsin Breast Cancer Data
Training Set Size = 369 Testing Set Size = 118

Additional experiments have been conducted comparing the performance BP and MSM on the WBCD and the Cleveland Heart Disease Data Set [Detr89]. The results of applying the methods to training and testing sets randomly extracted from the total points were averaged over many trials. These results are depicted in Figures 5 and 6. From these figures we draw the following conclusions:

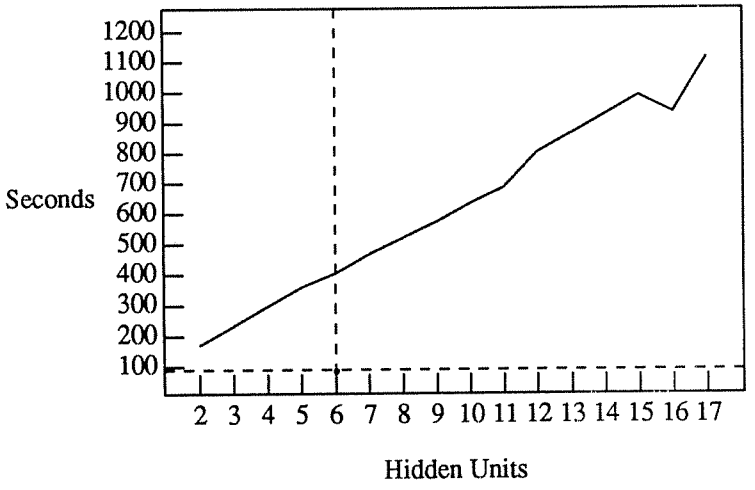
Training Set Correctness versus Hidden Units



Testing Set Correctness versus Hidden Units



Training Time versus Hidden Units



KEY:

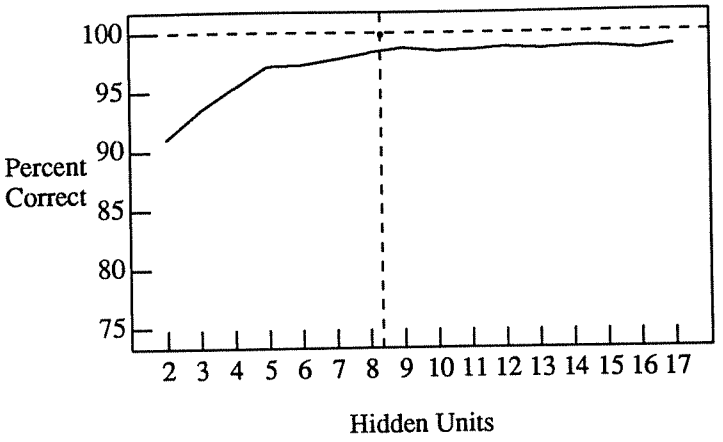
——— Back Propagation — Average of 10 runs. Data set contains 487 points.

------- Multisurface Method — Training Set = 67% Testing Set = 33%

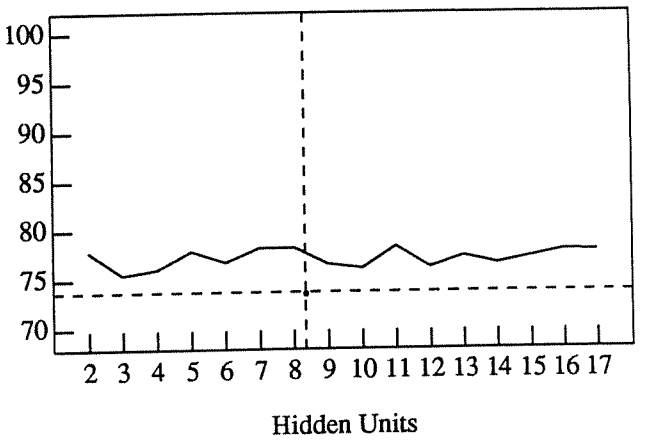
¦ Average number of hidden units for MSM

Fig 5: Results of MSM and Back Propagation on Wisconsin Breast Cancer Data

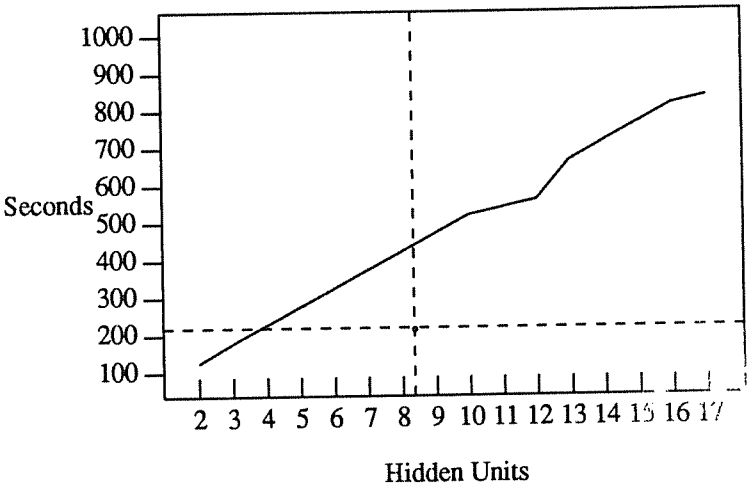
Training Set Correctness versus Hidden Units



Testing Set Correctness versus Hidden Units



Training Time versus Hidden Units



KEY:

——— Back Propagation

------- Multisurface Method

Average of 10 runs. Data set contains 297 points.

Training Set = 67% Testing Set = 33%

Average number of hidden units for MSM

Fig 6: Results of MSM and Back Propagation on Cleveland Heart Disease Data

(a) 100% correctness was always achieved by MSM on the training set but not by BP.

(b) On the testing sets, MSM achieved a correctness rates which are within 4% of the correctness rates of BP. The higher discrepancies occurred on the relatively noisy Cleveland Heart Disease Data set.

(c) The number of hidden units, which is determined automatically by MSM, is a good estimate for the number of hidden units required using BP to achieve minimal training time, and optimal training and testing set correctness.

(d) The time to train MSM is consistently much less than for BP. If we take into account that BP requires experimentation to determine the optimal values of learning parameters and the number of hidden units, the difference becomes more pronounced.

To sum up, MSM has the capability of quickly training a neural network and determining the optimal number of hidden units while maintaining 100% correctness on the training sets. These are important properties not possessed by BP.

REFERENCES

[Detr89] R. Detrano, A. Janosi, W. Steinbrunn, M. Pfisterer, J. Schmid, S. Sandhu, K. Guppy, S. Lee, V, Froelicher, "International Application of a New Probability Algorithm for the Diagnosis of Coronary Artery Disease", *American Journal of Cardiology*, 64, 304-310, 1989.

[Karm84] N. Karmarkar, "A New Polynomial Time Algorithm for Linear Programming", *Combinatorica*, 4, 1984, pp. 373-395.

[Khac79] L.G. Khachian, "A Polynomial Algorithm in Linear Programming", *Dokl. Akad. Nauk SSR*, 244(5), 1979, pp. 1093-1096.

[Mang65] O.L. Mangasarian, "Linear and Nonlinear Separation of Patterns by Linear Programming", *Operations Research*, 13, 1965, pp. 444-452.

[Mang68] O.L. Mangasarian, "Multisurface Method of Pattern Separation", *IEEE Transactions on Information Theory*, IT-14(6), 1968, pp. 801-807.

[Mang89] O.L. Mangasarian, R. Setiono, and W.H. Wolberg, "Pattern Recognition Via Linear Programming: Theory and Application to Medical Diagnosis", Computer Sciences Technical Report #878, University of Wisconsin-Madison, 1989, to appear, *Proceedings of SIAM Workshop on Large-Scale Numerical Optimization*, Cornell University, Ithaca, New Your, October 19-20, 1989.

[McCl87] J.L. McClelland and D.E. Rummelhart, *Explorations in Parallel Distributed Processing: A Handbook of Models, Programs, and Exercises*, MIT Press, Cambridge, MA, 1987.

[Rume86] D.E. Rumelhart, G.E. Hinton, and J.L. McClelland, "Learning Internal Representations", in D.E. Rumelhart & J.L. McClelland (Eds.) *Parallel Distributed Processing*, Vol. I, M.I.T. Press, Cambridge, Massachusetts,1986, pp. 318-362.

[Wolb88] W.H. Wolberg, M.A. Tanner and W.Y. Loh, "Diagnostic Schemes for Fine Needle Aspirates of Breast Masses", *Analytical and Quantitative Cytology and Histology*, 10, 1988, pp. 225-228.

[Wolb90] W.H. Wolberg and O. L. Mangasarian, "Multisurface Method of Pattern Separation Applied to Breast Cytology Diagnosis", submitted to *Proceedings of National Academy of Sciences*, USA.