Classification of Microcalcifications by Artificial Neural Networks

Masters Thesis: 424

Catholic University Nijmegen
Faculty of Mathematics and Computer Science
Section Experimental Computer Science for Technical Applications

Door:
Branko Johannus Michael van Abel

Mentor:
Dr. Th. E. Schouten

Preface

About two years ago I knocked on the door of Theo Schouten asking him for a subject for my masters thesis Computer Science. I went to him because the intricacies of artificial neural networks have always fascinated me and Schouten was known for giving programming assignments. The fact that he gives very practical assignments appealed to me because somehow along the way I run into some heavy delays while completing my masters and I was afraid a theoretical thesis would simply cost me too much time. Schouten confirmed my suspicions and assigned me the task of incorporating a batch language into the Convis neural network simulator. But my asking a subject for my thesis proved a bit premature because I first had to finish a number of obligated courses still missing on my resume.

When I came back a year later Convis had gone out of use, which again left me without a subject for my thesis. Schouten had at that time received the request from the radiology department at the Radboud University Hospital of Nijmegen of examaing a certain dataset containing mammograms. Maybe artificial neural networks could find some features with which they could classify the mammograms. Schouten tested the dataset and concluded that it was a very difficult dataset to learn. He asked me to perform further examination on the dataset to see whether it could be learned

For me the multidisciplinary aspect has been very interesting. The possibility to work with people outside my regular environment (the math and physics faculty) was new to me. Also the social relevance of the research made it very exiting. I hope that with this thesis I transformed that excitement into a well written, scientific exercise, which may be of some use to future research in the fields of mammography and neural networks.

Acknowledgements

First I would like to thank Theo ‘Keep on writing!’ Schouten for his guidance and valuable expertise on the subject and his patience with my sometimes tardy work. Nico Karsemeijer and Wouter Veldkamp I thank for providing the datasets and Wouter in particular for discussing the subject matter with me. Marcod saved me from a harmful ‘mouse-hand’ by forcing me to give him Ping-Pong lessons and giving the necessary mental relieve. By giving in to the habit of keeping the most valuable to the last I would like to thank my girlfriend Lonneke for conquering my bad moods and easing my mind. I think I’m not being pessimistic by saying that I definitely could not have done without her. So thanx to all,

Branko van Abel, August 1997

Contents

Chapter I: Introduction

With the advance of modern medicine the average human age seems to become higher and higher. Many diseases that were fatal in the past can now easily be treated and cured. Diseases like the plague and cholera were feared in the middle ages but nowadays lost their threat, that is, if you happen to live in the western world. But there are still diseases left which instil fear in peoples hearts. One of these disease is 'the big C', cancer. Most people know someone in their environment who suffers or died from the effects of cancer.

One of the most dangerous and most frequent forms of cancer is localised in the breast. In the Netherlands 30% of the women who have cancer, have breast-cancer. 20 percent of the women who die of cancer, die of this form of cancer. These figures are still increasing. Effectively this means 7000 new cases every year and 2000 fatalities caused by breast-cancer. Early treatment greatly reduces the risk of a fatal or disfiguring development of the disease. To ensure early treatment two screening programs have been started in 1975. These reduced the mortality rate among the screened population by 30%. The researchers use high resolution x-ray films of breasts. Albeit the performance has improved, it is still desirable to further reduce these figures. False alerts result in severe stress for the patient and her family while uncalled-for further investigation is carried out (happens in 2000 cases!).

At the start of this century Salomon, while taking X-ray films (hence forward named mammograms) of amputated breasts, noticed "small black spots" at the centre of **carcinoma**[1]. This was the first time microcalcifications in the breast have been noticed on a X-ray film. These calcium granules can indicate the presence of a malignant tumour, but they can also be milk leftovers. It would be a great advantage to determine whether one is dealing with a benign lesion or a malignant tumour. This way only excisional biopsies are performed in the cases already diagnosed as malignant. This would save in time and (often disfiguring) biopsies. However the classification method used should not allow so called false negative classifications to occur (in which case a malignant tumour is classified harmless). Through the years radiologists have improved the classification of these mammograms to a level of 50 percent, which still means that 50 percent of the people who undergo a biopsy have no malignant tumour.

At the academic Radboud hospital in Nijmegen a segmentation algorithm was invented and implemented which isolates microcalcifications on a mammogram. Although radiologists look mostly at clusters of calcifications, they were curious whether it is possible to identify malignant tumours using single calcifications. They are researching if this can be done using statistical and heuristic classifiers. A problem is the segmentation algorithm still being a little to rough and therefor allowing certain microcalcifications to enter the classifier which radiologists would not even consider classifying as malignant.

---

[1] Because the thesis covers several research fields the danger exists that incomprehensible jargon is used. For reasons of clarity and completeness added a thesaurus has been at the end of the thesis. All words which are explained in the thesaurus are written in bold face the first time they are used.

The branch technical approaches of the department of Computer Science of the University of Nijmegen thought this might maybe be accomplished by neural networks. Neural networks are computer paradigms which try to mimic the functionality of the brain on a very low level, the neuron level. A neural network can be trained by feeding it pattern response data. A learning set is used to train the network. At first it will produce random results but slowly it will match the correct response to a pattern. When a sufficiently high success level is reached a second set, the test set, is fed into the network to see whether it actually learned the correlations between the patterns and response. This way it is clear it not just 'memorised' the set.

Since the segmented pictures are relatively small (15x15 pixels) and test and learning sets are available it should not be a problem to let a neural network have a go at it. One clear advantage is that unlike the statistical and heuristic classifiers almost no medical knowledge is needed. This makes the research for this thesis in theory a multidisciplinary task but in practice it mainly requires knowledge of computer science. The central question posed in this thesis is thus whether it is possible to classify single microcalcifications using a neural network. It has been tried to come up with an answer to this question by first becoming familiarised by the field from which it arose and then try to experiment with the networks till there will be some results on which to base a conclusion. For the research project the microcalcifications together with their classification will be converted to a data for the backpropagation network. Half of them will be used for classification, half of them for testing. If the test set can be classified then the individual microcalcifications can be classified.

Summary of chapters

The first chapter deals with neural networks. It gives a brief history of neural networks. It describes their advantages and disadvantages compared to other ways of classifying data. It explains how the two neural networks used in this research project work. What is the theoretical background of the networks? What are the results so far using neural networks on visual input?

The second chapter gives more information on the specific neural network used, i.e. backpropagation. The origin of backpropagation will be revealed, and the mathematical paradigm, which is behind the neural network, will be explained in detail. The chapter also gives the advantages and disadvantages of backpropagation.

The medical background of this thesis is explained in the third chapter. It is tried to provide the reader with a basic knowledge of the anatomy and pathology of the breast. The origin of the microcalcifications is dealt with here and a taxonomy of the different calcifications is given. Furthermore it will shed some light on how mammograms are made and how the microcalcifications are isolated. Statistics of cancer disease will be given to emphasise the importance of early diagnoses of malignant tumours.

The following chapter describes the practical work performed for this thesis. It will be a report of the research project. Here it will be explained how the data is conceived and what kind of data is used. It elaborates on the kind of pre-processing was used on the data to improve the performances of the network. What parameters were used to tune the network to the dataset. The kind of application that was used to analyse the results will be shown as well as the kind of computer on which the networks have been implemented. Also the results of the other approaches will be depicted and compared to the results of the network approach.

Finally the last chapter contains the conclusions of this thesis. It analyses the results and tries to explain any failures. It gives an answer to the research question posed in the introduction and summarises this thesis. Furthermore it contains advice for future work on the subject.

Chapter II: Neural networks

People have always been amazed by the uncanny capabilities of the human brain. Scientists are amazed by its ability to gather the information provided by the senses, process it and react to it in a seemingly effortless manner. Attempts to reproduce by machines only small parts of this ability have proven to be very difficult. Trying to let a computer convert spoken language to words suffers under high error rates. For example trying to make a computer react to language is even more difficult because of the ambiguity of natural language. Scientists in the field of computer vision are having trouble letting the computer decide what information entered by its optical senses is important and what not. Therefor its not strange that scientists are looking at the human brain itself for answers. With the advance of neurofysiology this organ is lesser and lesser a black box. Physiologists already have an idea what the building block of the brain ,the neuron, looks like and how it functions. From this way up they try to unravel the secrets of the brain.

It is therefore not strange that scientists trying to copy the functionality of the brain start by copying the functionality of the neuron. The behaviour of this cell looks very much like a simple calculator. It receives input from other neurons, adds them together and if the total input reaches a threshold, it fires output to other neurons. This idea is easily formalised and implemented. If several of these artificial neurons are put together in a way analogous to the brain a network is created. Although the resemblance of this artificial network and the brain is only superficial, it seems to share some of the abilities of the brain. These networks can exhibit the ability to learn, generalise and abstract.

The learning capacities these networks are able to exhibit comprises the ability to learn from their environment. Many training algorithms have been developed to make a network approximate the correct responds to an input. Some of the algorithms are supervised, i.e. they have to provide a set of correct pattern-responds combinations. Others networks are unsupervised, in which case the network tries to learn to discern certain similarities in the patterns by itself. The advantage of supervised learning is that it gives the operator of the network more control of what the network learns. Unsupervised learning on the other hand loosens the restrains on the network, giving it more room to find by itself certain characteristics in the input.

The ability to recognise certain characteristics in the input patterns is called abstraction. Especially in fields where theory is limited, abstraction by a computer is a valuable asset. One can use a network to see if there is certain information to be gathered in a range of inputs and if it finds some then one can start to come up with a theory.

Generalisation is another important feature of neural networks. Conventional programs are often too specific. They do not seem to be able to recognise patterns, which have been distorted by noise. Since we do not live in a perfect world, such distortion nearly always seems to occur. The ability to overcome slight variations in input and then provide the correct responds is therefore a very important ability, which is very convenient when for instance trying to read hand-written texts.

Concluding one can say that, although artificial neural networks are far from being exact copies of their biological counterparts, they do try to simulate the operation of the human nervous system. In doing so they exhibit features ascribed to human thinking like learning, abstraction and generalisation. In this thesis I will try to make use of those features in the hope the neural networks can spot correlations in the data which are very hard to see with the naked eye.

Summary of sections

The first section gives a brief run-through of the recent and turbulent history of neural networks. Then the biological prototype will be examined to provide an example for its artificial counterpart in the third section. The training of the artificial neural nets will be the next subject and the difference between supervised and unsupervised training. The chapter will be concluded with a description of the abilities and disabilities of neural networks as opposed to conventional computer programs.

1.History of the field of artificial neural networks.

The idea of neural networks first emerged in 1949 with the learning law of D.O. Hebb [Hebb 1961]. The neurophysiology, neuroanotomy and psychology were developing models of learning and Hebb's law was one of them. Scientists believed that networks based on this law could be capable of learning. In the following decades the anatomical and psychological knowledge was combined to create the first neural networks. First they were build on electronic circuits, but later they were simulated on computers. The latter, being a much more flexible approach, is still the most commonly used today. Early results with these single layer networks were so promising that a huge amount of money and resources were allocated to research these 'brain like' computers. The expectations were extremely optimistic. It was believed that to create an artificial brain was only to build a large enough network. It was this overconfidence in the capabilities of neural networks, which led to their downfall.

One of the most renowned researchers in the new field of artificial intelligence Marvin Minsky noticed among others that neural networks could not solve some simple problems. This was strange because problems very similar to the ones that failed were easily solved. The suspicion arose that neural networks had a kind of blind spot in their problem solving. Minsky attacked the problem with mathematical techniques and came up with theorems regarding neural networks. In his book Perceptrons Minsky [1969] not only showed but also mathematically proved that single-layer networks were unable to simulate a whole class of functions including a simple exclusive or gate. These linear separable functions constituted such a huge subclass of all possible functions that the use of single-layer networks as a general problem solver is impossible. Minsky added to it that since a single level network could emulate multiple level the restriction also applied to the multiple level networks.

Minsky had a severe effect on the field of artificial neural networks. The facts that Minsky had huge prestige in the field of artificial neural networks and the impregnability of the mathematical prove changed the euphoria into a deep depression. Researchers left the field and governments stopped their funding. The research of artificial neural networks entered its dark ages. They would last for nearly two decades.

Only a few researchers stayed in the field working with the few resources they had. The downfall made it difficult to find publishers and to cooperate with other scientists because articles would be denied. But thanks to the painstaking effort of a few researchers such as Kohonen and Grossberg the field rose from its ashes as a phoenix reborn. A strong theoretical fundament was constructed on which the new multi-layer networks were build. Artificial networks started to solve problems Minsky thought impossible. Minsky book proved to be to pessimistic and slowly but surely the field regained its momentum in the eighties and now looks healthier then ever with multiple commercial applications as natural speech (NetTalk) and recognition of printed (OCR) or written text.

2.The human neurons

The concept of artificial neural networks (hence forward just neural networks) originates from the brain and therefore it is logical that to study the neural networks one first looks at their biological counterparts. The nervous system consists of neurons. These cells serve as little processors which are able to receive, process and transmit further on the signals they receive from other neurons. They do this by using an ionic fluid.

The way the 'data processing' works in detail can be shown by means of Figure II-1. A neuron consists roughly of three parts. The cell body, the axons and the dendrites. The cell body is the main processor, the axons serve to distribute the signals, and the dendrites serve as roots, connecting the neuron to other neurons. The connection point is called the synapse. First the signals from others neurons are received at the synapses. Here the input from other neurons is summed, some excite the cell, and some inhibit the cell. The signal follows the dendrite where they reach the cell body. Here all the inputs from the dendrites are gathered. If the accumulation of all inputs reaches a certain threshold, the cell fires and sends a signal by way of its axon to the other neurons. The whole operation takes place in a purely asynchronous manner. This is a somewhat simplified characterisation of a neuron, but useful enough to serve as an example for the artificial neuron.

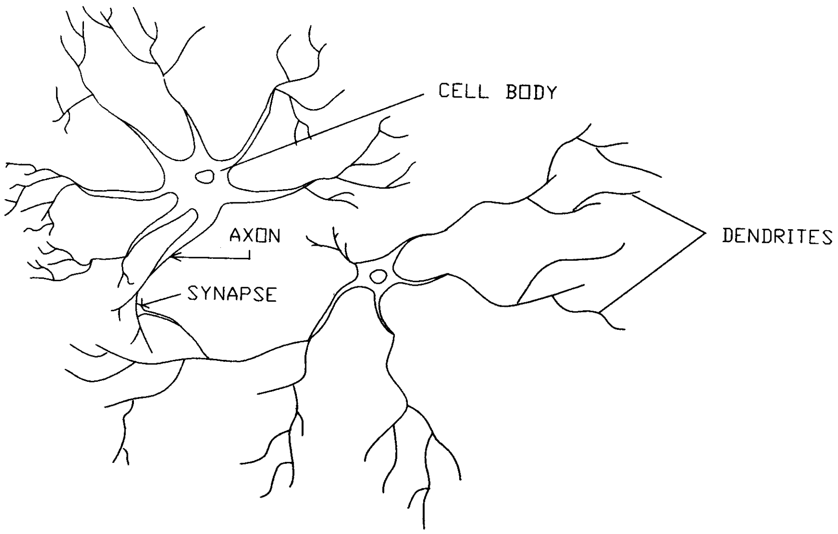

Figure II-1: Human neuron

The whole operation of receiving signals and firing takes place in hundreds of a second. The question arises why we did not already simulate a brain on a computer; after all contemporary microprocessors work at speeds of nanoseconds. The answer lies in the complexity of the nervous system. The nervous system consists out of 10^{11} neurons; the average neuron has roughly 10000 connections with other neurons making it 10^{15} interconnections and they all work in parallel. What it lacks in speed it makes up in numbers and parallelism.

Still scientist wonder how something that consists of cells which function very similar to a simple calculator can be the basis for human cognition. Modern science made dualism an unscientific idea and if there is not an immortal soul where does the consciousness come from? The answer seems to lie it the brain but how this is accomplished is still the subject of heavy scientific and philosophical debate. Fortunately the subject of consciousness falls outside of the scope of this thesis but further reading on this subject can among others be found in Dennett's book [1991].

3.The artificial neuron

As mentioned earlier the artificial neuron tries to simulate the functionality of the biological model. A neuron is connected to other neurons from which it receives its input. Each connection has a

weight, which excites or inhibits the input similar to the synaptic strength. The output of the summation function is applied to a activation function **F** producing the **OUT** value. This can be formalised as follows [Wassermann 1989]:
A neuron has connections with input $i_1, i_2, ..., i_n$, these inputs form input vector **I**. All connection have weights $w_1, w_2, ..., w_n$ and these form the weight vector. The vectors can then be multiplied producing the NET output. This multiplication is the algebraic function **NET** = $\mathbf{IW}^t$, as depicted in Figure II-2.

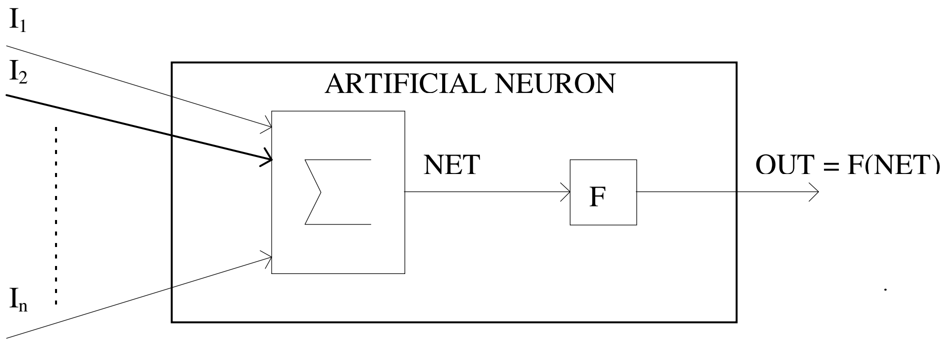

Figure II-2: Artificial neuron + activation function

After the NET-output is calculated, an activation function is applied to it. The simplest form for this function is a simple linear threshold function, which fires if the net output is above a certain constant value. The problem arises when the input has great variation in size. Choosing a too small value for the threshold makes the network ignore differences in high values, taking a to big threshold ignores the differences in low values. To avoid this problem one can to transform this into a non-linear function which more looks like the operation of the biological neuron. This can be done by first 'squashing' the net output before applying it to the threshold function. The function most used is the logistic or sigmoidal (S-shaped) function (See Figure II-3)

$$f(x) = 1/(1+ e^{-NET})$$

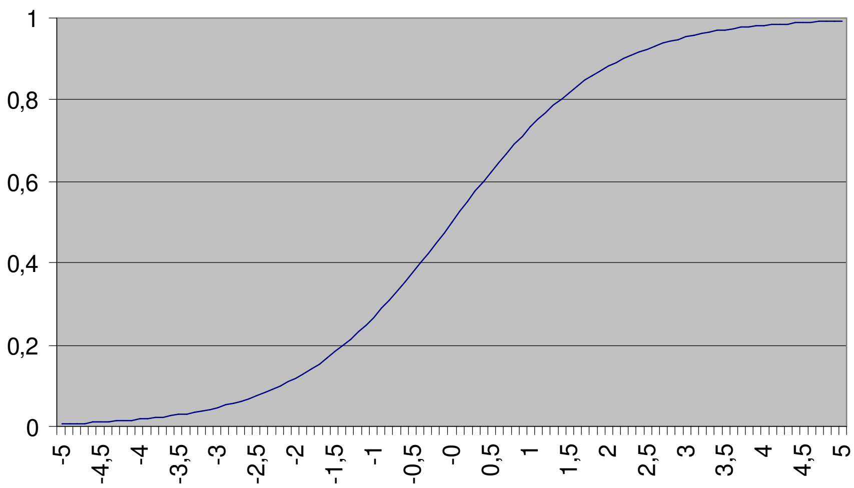

Figure II-3: Sigmoidal logistic function.

This function has as an advantage that the value applied to the threshold function never exceeds a certain value (for this function the value 1). It solves [Grossberg 1973] the noise saturation-dilemma, i.e. how can a network at the same time work with small and large signals? The form of the function solves this problem. For small values (NET near zero) the slope of the input/output curve is steep,

producing high gain. When the value of signals increases the steepness decreases, which means the gain decreases. This way large signals can be processed by the network without saturation, while small signals are accepted without weakening the network.
Another squashing function is the hyperbolic tangent. This function has the same shape and characteristics but is also bipolar which seems to work better for some networks. It has a output of 0 when the NET output is 0 and it ranges from -1 to 1 (See Figure II-4).

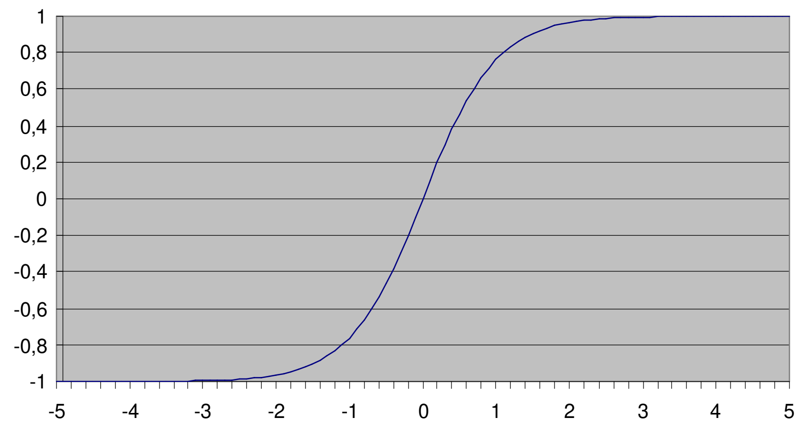

Figure II-4: Hyperbolic tangents

Notice that the activation functions mentioned are only approximations of the actual process taken place in the neuron. But nevertheless these functions give the artificial neural network the possibility to mimic the capabilities of the brain.

4.Training

Networks need to be trained in order to assimilate the knowledge they are presumed to learn. D.O. Hebb tried to mimic the operation of the synapses in the brain. For this synaptic strength of the connection between two neurons is increased if both neurons are activated. This way the phenomenon of learning through repetition can be explained. The rules he came up with became the fundament of network training.

In Hebbian training the value of the weight of a connection is increased by the product of the excitation of the neurons it connects. Symbolically:

$$w_{i,j}(n+1) = w_{i,j}(n) + \alpha \cdot OUT_i \cdot OUT_j$$

$w_{i,j}(n)$ = the value of the weight connecting neuron i to neuron j prior to adjustment.
$w_{i,j}(n+1)$ = the value of the weight connecting neuron i to neuron j after adjustment.
α = learning rate coefficient
OUT_i = the output of neuron i and input to neuron j
OUT_j = the output of neuron j

Many networks have been constructed with Hebbian training, but nowadays more effective learning methods are used. These methods will be described in the next chapter.

Supervised and Unsupervised learning

When applying supervised learning, there is already some information available about the proper reaction to a certain input. This way a training vector can be constructed, a pair consisting of the input vector and the desired output vector. This is called a training pair. The learning process goes as follows: the input vector is offered to the network and a result vector is calculated. The result vector is then compared to the desired output and the difference between the result vector and desired vector is calculated (error value). The difference is fed back into the network to use it to adjust the weights to minimise the error. If all goes well, repeatedly applying a batch of training pairs will minimise the error and the network will converge. The network has then 'learned' the training set. The backpropagation network described later on is an example of supervised learning.

Scientists claim artificial neural networks are a simulation of the real thing. Supervised learning makes this claim implausible, in the brain usually no desired output exists and still the brain seems to be able to learn a response to a certain stimulus. How does a child learn, how does the brain organise itself? Apparently most learning done in the brain is unsupervised learning. Kohonen [1988] and others came up with an algorithm to simulate this unsupervised learning where no ideal response exists. Unsupervised learning requires only a set of input vectors. These input vectors are repeatedly applied to the network where the training algorithm adjusts the weights to produces a consistent output vector, i.e. if one applies the same input vector twice, or a very similar vector, the same output vector will be produced. This way the training process will divide the training set into classes, where each class consists out a group of vectors which have very similar statistical properties. All vectors in a class will produce the same output vector but there is no way of knowing what that output vector exactly is. Therefore, the output vectors of the different classes are in the Kohonen implementation mapped to a two-dimensional map of neurons. Every time a input vector is applied to the network a corresponding neuron on the map lights up, signifying the class found.

5.Single layer versus multiple layer.

Neural networks consist of layers. Each layer is connected to the next. The first networks, named perceptrons, are single layer networks. They had what appeared to be two layers, but the first layer only worked as a gatherer for all the input and did no do any computation of its own. In neural computing one in general does not count the first layer. To avoid ambiguity it is best to mention the number of hidden layers, i.e. all the layers except the output layer and the input layer. Which means perceptrons have no hidden layer. Figure II-5 shows a 1-hidden-layer network. In a fully connected network every neuron of a layer is connected to every neuron in the following layer.

When perceptrons failed to solve simple problems like xor-gates, it was also assumed that multi layer networks would not offer a solution. As long as one uses linear activation functions every multi layer networks can be simulated on a single layer network. Recall the vector multiplication, which was processed in the network. Suppose we have a two-layer network (See Figure II-1).

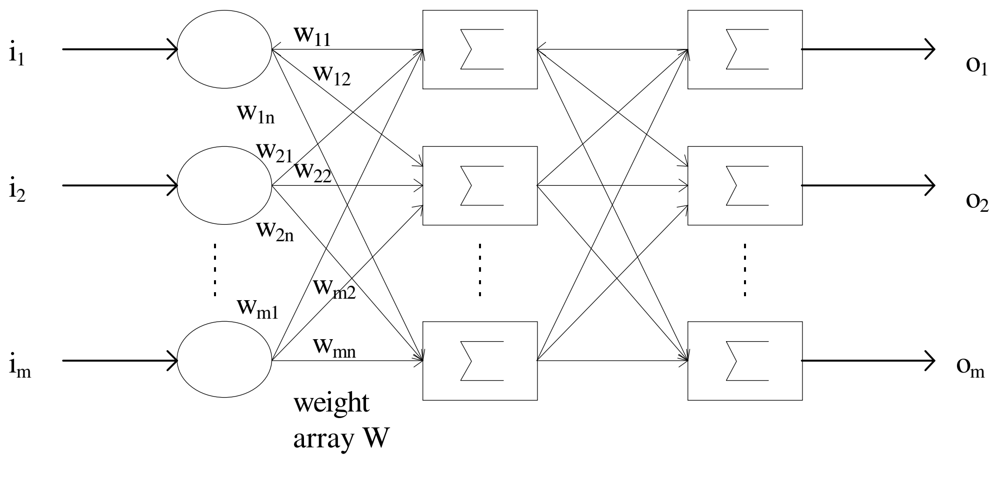

Figure II-5: One hidden layer network

An input vector **I** enters the network. It is then multiplicated with the weight vector $\mathbf{W}_1$ of the first layer. Next it is multiplicated with the weight vector of the second layer $\mathbf{W}_2$ producing the output vector **O**. Since the activation function is linear this can be formalised as the algebraic function:

$$(IW_1)W_2 = O$$

Matrix multiplication is associative so it also means:

$$I(W_1W_2) = 0$$

Now a third vector $\mathbf{W_3}$ can be made with $\mathbf{W_3 = W_1W_2}$ given the final function:

$$IW_3 = 0$$

So the two-layer network can be simulated with a single layer network having the matrix product of the weight vectors of the two layer network. This proof can be extended with induction to n-layer networks so anything which can not be solved with a single layer network can not be solved by a multi-layer network. But as mentioned this only works with linear activation functions and not with non-linear activation functions, making it possible that a multi layer network may succeed where a single layer network fails.

6.What neural networks can (not) do.

The advantage artificial neural networks (ANN) offer as opposed to heuristical approaches is clear. ANN's can work in the absence of a valid, consistent theoretical background. If there is the suspicion of a relation between a set of inputs and a set of outputs, let an ANN crunch on it for a while and maybe the relation is found. All that is needed is a training set and a test set. By first applying the training set to learn the relation and then applying the test set to verify whether the network really learned something. The distinction between supervised and unsupervised learning provides a certain degree of autonomy of the learning algorithm. With supervised learning it is still possible to steer the network somewhat in the right direction, whereas unsupervised learning is almost beyond control. This autonomy is as well the premium advantage as major disadvantage of the neural network. It makes it very flexible and able to coop with noisy input but it forces the user to give up control.

It is very difficult to see if a network has learned the correct relation. An amusing anecdote is provided by a research by the ministry of defence (a very large part of the funding of the field is supplied by defence). They tried training a neural network to spot tanks. Pictures were taken of tanks in the field, tanks in the woods, tanks with camouflage etc. and pictures of fields, woods without tanks. They trained the network with the pictures and after a while it converged. The network was able to say, with a certain degree of error, if the picture contained a tank, yes or no. But when they fed the test set to the network it performed poorly. What happened? The simple solution soon emerged that the pictures of the training set with and without tanks were taken on different days. One day it was clouded and the next day had a clear sky. Ergo the network learned itself to discriminate cloudy days from clear days.

The anecdote is a very good example of one problem with neural networks. It is difficult to see if the network learns the correct relation. It also shows that although one can't force the network, one can manipulate the learning process by letting the training set focus on the relation to be learned. The set should as much as possible be free from accidental relations and be as random as possible. In general it can be said that the bigger the training set the better. Unfortunately learning time increases with the size of the set, and large sets are expensive. So a compromise has to be chosen between set big enough to exclude accidental relations and small enough to keep the leaning time acceptable.

One has to keep in mind that ANN's are not the new improved version of conventional computer programs. It is more realistic to consider them a powerful new class of computer programs designed to solve specific problems, especially imitating human capacities as learning and abstraction. There are a whole class of problems in which they consume too much time and space. Simple precise calculation is one of the tasks better handled by conventional programs; manipulation of symbols and memory should also be left to conventional programs. Artificial neural networks are not the panacea they were thought to be, but can be a very powerful tool when applied to their own class of problems.

Chapter III: Backpropagation

As mentioned in the introduction the field of artificial neural networks entered a period of low interest after the publication of Minsky's book. One of the first networks being able to cope with the problems mentioned in the book was backpropagation. It was one of the networks, which succeeded in bringing the field of neural networks back into the spotlight again. Although having several drawbacks as network paralysis and local minima its is still one of the most frequently used today.

It is difficult to attribute a father to the backpropagation algorithm. Rumelhart, Hinton and Williams [1986] seemed to be the first to give a good description of the algorithm. After the article was published Parker [1982] showed that he invented the theory before them. To make it even worse Werbos [1974] had already described it way before him. Reinventing the wheel is even in this age of information not uncommon for science in general. This is especially the case for artificial neural networks due to the overlap of multiple fields. To be truly universal in the field one has to keep track of books and articles scattered over many fields of science.

Backprop is short for "backpropagation of error". Backpropagation refers to a method of computing the error gradient of a feedforward network, a straightforward but elegant application of the chain rule of elementary calculus [Werbos 1994]. Implementing the theory this leads to a feedforward network which is trained by the backpropagation of the error.

Summary of sections

First section will explain in detail the configuration of the backpropagation network, show how the neurons function and how they are connected. It will be succeeded by a section dealing with the training and operation of backpropagation. The mathematical functions will be given symbolically as well as the principles behind them in plain words. Also important training parameters as momentum and learning rate will be presented. The chapter will be completed with a roundup of the problems connected to implementation of the backpropagation paradigm and a number of examples of successful applications.

1.Design

The design of backpropagation differs very little from the original architecture of the first artificial neural networks. Backpropagation consists of a number of neurons grouped by layer. All neurons in one layer are connected to all neurons in the next layer (non-fully connected networks also exist but are not so common). The first layer is only used as an information gatherer where the input vectors are fed to the network. The last layer contains the output vector. In between may be one or several so called hidden layers. The connections between the layer use weights to inhibit or excite the signal of the firing neuron (See Figure III-1).

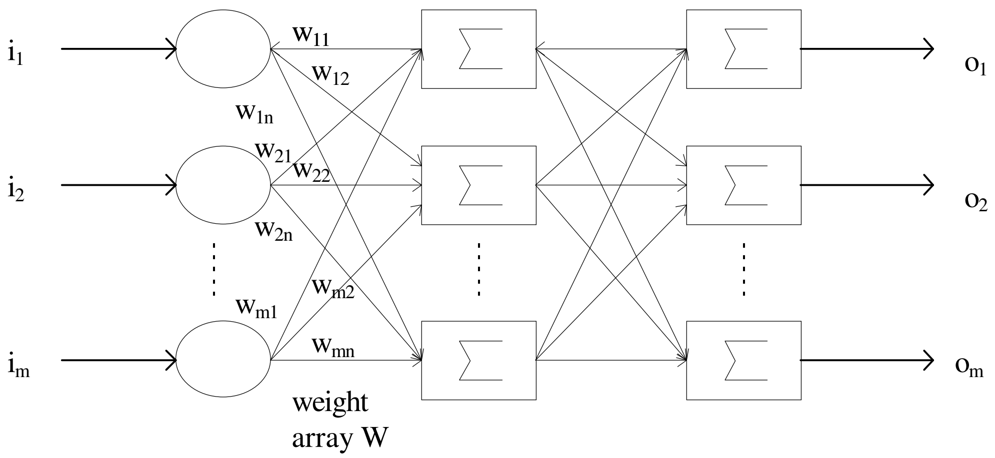

Figure III-1: Backpropagation network

All neurons except those in the first layer operate in the same way. First all signals from the neurons in the previous layer are gathered. These signals are excited or inhibited by the weight of the connection. All signals are summed to create the NET-value. This value is then applied to a squashing function (usually a sigmoidal function) to create the OUT-value. The OUT-value can then be send to the next layer or be part of the output vector. The usage of a sigmoidal squashing function is not obligatory but is commonly used because it provides automatic gain control. It should be differentiable to allow the error being propagated back into the network. Using this squashing function should cause large values not to saturate the network and small values not to weaken the network. Because this squashing function is non-linear more than one layer can be used. If just linear functions were used then a single layer network with all its drawbacks could emulate a multi-layer (See II-5 Single layer versus multiple layer. page 8).

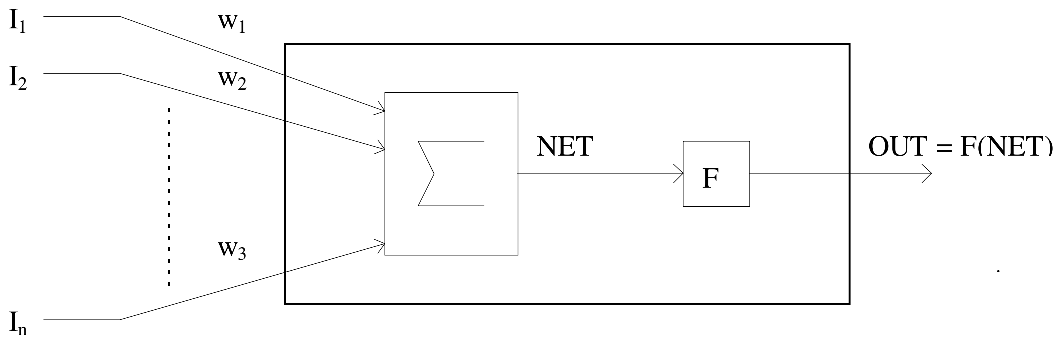

Figure III-2: Neuron in backpropagation network

$$NET = i_1 \cdot w_1 + \ldots + i_n \cdot w_n = \Sigma\, i_k . w_k + threshold$$
$$OUT = F(NET)$$
$$F(x) = 1 / (1 + e^{-x})$$

2.Network training

The training of backpropagation is supervised, i.e. to make the network learn a certain input/output function a dataset has to be at hand with a number of input vector combined with the correct output vector. The combination of an input vector and an output vector is called a training pair and is used to train and test the network.

To use backpropagation to classify a certain input set, one first has to prepare the network. This preparation involves two stages, training and testing. For this purpose two datasets are needed with different training pairs. Usually this is done by taking the vectors from the dataset and divide them in turn over the training dataset and the testing dataset. The training set is used to make the network learn the desired input/output relation. If the network is able to produce the correct output to a certain input within a certain error margin for all vectors in the training set then the training process ends. If the network is able to provide the correct output vector for each input vector in the test set then the network apparently approximates the I/O-function. This does not always have to be the case. It is possible that after the convergence of the training algorithm the network scores very poorly on the test set. This can occur for numerous reasons, for instance that the test and training set are not randomly enough mixed or the network somehow stored all training-pairs from the training set into the network without extracting the I/O-function.

The training algorithm consists of two passes, the forward pass and the reverse pass. In the forward pass the input vector is fed into the network and an output vector is calculated. In the reverse pass the network determines the difference between the output vector and the target vector to come up with an error value. This error value is then propagated back into the network to adjust the weights to reduce the error value. Both passes are performed for each training pair. Testing a network is the same as training but without the second pass. The training algorithm is as follows:

1. **Select the next training pair from the training set; apply the input vector to the network input.**
2. **Calculate the output of the network**
3. **Calculate the error between the network output and the desired output (the target vector from the training pair)**
4. **Adjust the weights of the network in a way that minimises the error.**
5. **Repeat steps 1 to 4 for each vector in the training set until the error for the entire set is acceptably low.**

Steps 1 and 2 constitute the forward pass. The training pair is composed from an input vector **I** ($\mathbf{i_1}$ to $\mathbf{i_n}$) and a target vector **T** ($\mathbf{t_1}$ to $\mathbf{t_m}$). The input vector is applied to the network by giving the **n** 'neurons' in the input layer input $\mathbf{i_1}$ to $\mathbf{i_n}$. Notice that these neurons do not perform any computation at all but gather and distribute the input. Each neuron in the input layer sends its signal to all the neurons in the next layer. The operation of the neurons in the following layers is identical. They receive input $\mathbf{i_1, ..., i_n}$ from the neurons of the preceding layer. These inputs are summated producing the **NET** value and then squashed to create the output value **OUT**. The **OUT** value is part of the output vector **O** of this layer.

If it is the last layer of the network then the output vector is the result from the network, to be used for classification or to compute the error value. Otherwise the **OUT** value is sent over the connections to the neurons in the next layer. Before it can be used as input for the neurons in the next layer, it is multiplied with the weight value of the connection ($\mathbf{w_{i,j}}$ of neuron **i** from layer **n** neuron **j** from layer **n+1**). This operation can be expressed in matrix multiplication as:

$$I_{n+1} = O_n W \ \ i.e. \ \forall_{k,l} [\, i_{l,x} = \Sigma\, w_{k,l} \cdot o_k]$$

$\mathbf{I_{n+1}}$ = input matrix for layer **n+1**
$\mathbf{O_x}$ = output matrix of layer **n**
$\mathbf{w_{k,l}}$ = weight of connection between neuron **k** in layer **n** and neuron **l** in layer **n+1**
$\mathbf{i_{l,x}}$ = input **x** of neuron **l**

The reverse pass in step 3 and 4 comprises the learning capacity of backpropagation. Here the error value is calculated and adjustment of the weights takes place. Adjusting the weights is not the same for all layers. Only the last layer has a target vector (from the training pair) to compute an error value. The target vector and error value of the preceding layer have to be calculated by propagating the error back into the network.

The adjustment of the weights involves the so called delta rule. Say we have an output value **O** and a target value **T**. Then δ is the difference between **O** and **T**.

$$\delta = (T - O)$$

If δ = o then the neuron produces the desired output and no adjustment of the weights is necessary. If delta is not 0 then the weight is adjusted by adding delta multiplied by the input value ($\mathbf{x_i}$). The product is then again multiplied by a constant value indicating the learning rate (η, usually varying between 0.1 and 1).

$$\Delta_i = \eta \cdot \delta \cdot x_i$$
$$w_i(n+1) = w_i(n) + \Delta_i$$

$\mathbf{\Delta_i}$	= adjustment for the i^{th} input $\mathbf{x_i}$
$\mathbf{w_i(n)}$	= weight value i before adjustment
$\mathbf{w_i(n+1)}$	= weight value i after adjustment

To use this rule in backpropagation imagine the following setup. We have neuron **j** in the output layer and neuron **i** in the preceding (hidden) layer. We want to adjust the weight of the connection between the two neurons (**w[i,j]**). First we compute the difference between output value **O(i)** and target value **T(i)** to produce the ERROR signal. Then we have to compensate for the effect of the squashing function by multiplying the difference with the derivative of the function. For the standard sigmoidal function (Figure II-3) a simple derivative exists [O(I - O)]

$$\delta = O(1 - O) \cdot (T - O)$$

This delta is then multiplied with the OUT value of neuron **i** (the input value of neuron **j** before applying the weight change). This product is exactly the difference between the value of the weight now and the value it should have been to produce an output equal to the target value. A multiplication with η gives the user a parameter for manipulating the amount of change per input. In symbols:

$$\Delta w_{i,j} = \eta \cdot \delta_{i,j} \, OUT_i$$
$$w_{i,j}(n+1) = w_{i,j}(n) + \Delta w_{i,j}$$

$\mathbf{w_{i,j}(n)}$ = the weight of the connection between neuron **j** in the output layer and neuron **i** in the preceding layer before the change.
$\mathbf{w_{i,j}(n+1)}$ = the weight of the connection between neuron **j** in the output layer and neuron **i** in the preceding layer after the change.
$\mathbf{\delta_{i,j}}$ = the calculated δ for the weight between neurons **i** and **j**
$\mathbf{OUT_i}$ = the output value of neuron **i** in the preceding layer

Changing the weights for hidden layers.

The calculation of the weights between hidden layers is almost the same as the one written above. Except one variable is missing, the value for **T**, the target value. So a new way has to be found to calculate the delta value. This is where the backpropagation of the error, which gives the algorithm its name, takes place. After the weight adjustment of the connections between the output layer and the last hidden layer, all weights between these layers have their deltas calculated. Looking at one neuron in the last hidden layer, the **OUT** value is known (calculated in step 2) as well as the deltas of all the connections to the last layer. So by multiplying the deltas with the original OUT value, a value can be constructed what the OUT value should have been. This gives the new value for delta for a connection between neurons in the hidden layers:

$$\delta_{i,j} = OUT_i \, (1 - OUT_i) \, (\Sigma \, \delta_{j,k} \cdot w_{i,j})$$

This way the weights of the connections between the hidden layers just preceding the output layer can be calculated. These two hidden layers can then be seen as a hidden layer and the output layer, so the process can be repeated for the layer preceding these two layers. Working backwards through the network one can now adjust all the weights between layers.

Adding momentum

Rumelhart, Hinton and Williams [1986] improved training time and network stability by giving neurons a kind of memory what the last weight adjustment was. If the change of weights moved in a certain direction and suddenly somehow changed to an opposite direction then the momentum of the movement would restrain the opposite change and somewhat stabilise network adjustment. This can be done by adding the last weight change multiplied by the momentum (α) to the current weight change. In symbols:

$$\Delta_i(n+1) = \eta \cdot \delta_i \cdot OUT_i + \alpha \cdot \Delta_i(n))$$
$$w_i(n+1) = w_i(n) + \Delta_i(n+1)$$

This value of momentum α is usually set to 0.9.

A visualisation of the error surface looks like a terrain with hill and valleys. The lowest place in the lowest valley represents the optimal (lowest) error value. The network tries to find this place and momentum helps the network to follow the descend of the hills more gradually without running across the surface from one side to the other. Momentum usually improves learning time but can work contraproductive for some problems.

3.Problems

Backpropagation solved a lot of problems perceptrons could not, but backpropagation also has its own drawbacks. Very long training times are not uncommon, probably as a result of non optimal training steps. The network may fail to train all together; a problem brought about by local minima or network paralysis. Here follow these problems in more detail.

Network paralysis

After a certain amount of training the weights in a network can be adjusted to large values. To allow the network to process small as well as large values a sigmoidal squashing function is used. For large values this function calculates very small changes. Small changes cause small adjustments of the weights, so small that the network does not change anymore. Ergo the network does not learn anymore. The problem can be avoided by using smaller values for the learning rate (η), also named step size, but this results in longer training times.

Local minima

As mentioned earlier visualisation of the error surface gives a multidimensional terrain with hills and valleys. The ultimate goal is to reach the lowest point on the terrain. Imagine the current error state as a ball rolling over the surface. Backpropagation functions as a kind of gravity making the ball roll down the hills. The concept of momentum is easy to transpose to this analogy as the real momentum of the ball. Problem is that the ball will find the lowest point in the current valley but next to the valley may be an even deeper valley. The only way to reach it is by climbing up the hills but this is just the opposite effect of the way backpropagation (as well as gravity) works. An idea would be to give the ball more speed to make rolls over hills to find the next valley but then it is more difficult to make the ball stop. In network terms one can increase the momentum and lower the learning rate (step size) but it will take longer to stabilise the network (or the network will not stabilise at all). A theory for artificial neural networks which may have a better chance to find the global minimum has been developed [Wasserman 1988a] but will not be used in this thesis for lack of a fast implementation.

Non optimal step size and network size.

According to the proof of Rumelhart, Hinton and Williams (1986) backpropagation will always converge but it may take an infinitely small step size (η). This means it will take infinite time to converge, which makes it a somewhat impractical tool. Choosing the right step size is difficult and can only be determined by experience. Choose too low and the network will take too much training time, too high and network paralysis or continuous instability can occur which makes it impossible for the network to ever reach an acceptable error value.

Basically the same problem is the size of the network. How many neurons do you want to have in the hidden layer, how many layers do you want? Larger networks have greater storage capacity, which can mean they can learn more, but also can mean that they just store the whole training set in the weights. Larger networks have more weights and more neurons so longer training times. Finding the right size can be an arduous and boring task.

4.Practice

Backpropagation is used in a number of fields such as optical character recognition by NEC, converting English text into intelligible speech (NetTalk, [Sejnowski and Rosenberg 1987]) and image compression. At the university Nijmegen it has been tested on a number of applications like written script recognition with varying success. The mathematical paradigm allows the computation to be divided over a number of computers or processors, making it easy to build a parallel implementation. This is not surprising since it was based on the highly parallel brain architecture. Such a parallel implementation severely cuts down on training time. Applications show that is a powerful tool for recognising patterns in noisy input. However further research to increase its effectiveness and overcome its problems is advisable.

Chapter IV: Medical background

Artificial neural networks do not interpret the data that they process. For the network the input vector is nothing more than a number of values. They receive the values in the first layer and produce outputs according to the i/o-function that they perform. Whether the inputs are images from stars from a space telescope or (in this case) images from microcalcifications from x-ray machines is not important. The network just needs abstract values and sometime (supervised learning) correct responses to those values.

One could argue that for a thesis in artificial networks no information or theory about the dataset is necessary. One of the important features of artificial neural networks is that they can operate in absence of such a theory. However I do feel that some information about the background of the dataset is necessary. First of all it adds some spice to the flavour of an otherwise purely technical thesis. More important it shows why such a research project is needed. Not only for collecting information about the usage and applications of neural network, but also to try to help formulate or make applications for the area the dataset is from.

The need for extra research in the field of radiology for breast cancer is eminent. American research shows dramatic figures. One out of eight women gets breast cancer. In the States every three minutes a woman is diagnosed with breast cancer, every 11 minutes a woman dies from breast cancer. A total of 46,000 women and 300 men will die of breast cancer this year. Dutch statistics show that 30 % of female cancer is located in the breast and 20 % of cancer fatalities is by breast cancer [Doyle 1995] and the figures are increasing. Two experimental screening programs in Nijmegen and Utrecht were initiated in 1975 to decrease these figures and showed a 30 % decrease in fatalities from breast cancer for the screened population [Martens et all 1990].

Summary of sections

Starting with the basics is a section about the anatomical features of the (female) breast. The main function and composition will be explained. Then a section revealing the facts about breast cancer, what is it and what kinds are there? To treat the effects of breast cancer it first will have to be detected. The next section will elaborate on methods of detection, followed by an in depth description of the detection method of which this research project tries to contribute, the radiological classification of microcalcifications. For reason of completeness an extra section is added about the treatment of breast cancer.

1.Anatomy of the breast

The breast is an **exocrine gland**[2] that produces milk during the period of lactation. The main function of the breast is to produce milk. Each breast consists of 15 to 20 sections called lobes, which look like the coloured leaves of a daisy. Each of those leaves contains many smaller lobules. These lobules end in numerous tiny milk-producing bulbs. Small thin tubes called ducts connect these lobes, lobules and bulbs. The ducts all lead to the nipple which is the centre of the dark area on the breast. This dark area is called the areola. The remaining spaces in the breast are filled with fat. The pectoral muscles, which cover the ribs, are underneath the breasts but are no part of it. From the muscles to the skin of the breast lie tough bands of tissue called **ligaments**, which support the breast. These are the origin of skin retraction as a result of the **tumour** deep inside the breast.

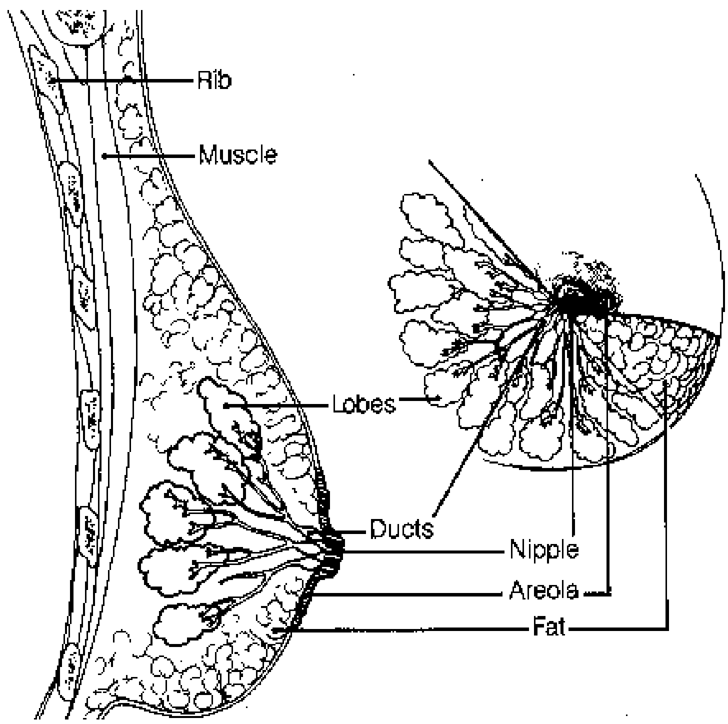

Figure IV-1: Anatomy of the breast.

2.Cancer in the breast.

Cancer is not one specific disease but a category of diseases which all have the same effect, abnormal cell growth and destruction of body tissue. Normal cell growth causes the cells, which make up the body tissue, to grow, divide and replace themselves. Sometimes cells lose the ability to control their growth. They grow too fast, in a disorderly fashion and create too much tissue and tumours.

Tumours can be divided in two main categories: benign and malignant tumours. Benign tumours grow but do not pass **histological** borders, which means they do not spread through the body. Although they can grow to huge sizes, they are in general not fatal and normally they can be removed. Benign tumours are not cancer. Examples are **cysts** and **fibroadenoma**.

Malignant tumours can pass histological borders and infiltrate surrounding tissue. They spread around the body using blood- and lymphvessels causing **metastasis**. Metastasis is the spreading of the disease to other parts of the body. For instance the malignant tumour in the breast can be removed by a mastectomy, but the cancer can have spread to other body parts.

If a malignant tumour is found early, it can be still in a state in which it did not yet pass histological borders. This is called **In Situ** malignancy or non- invasive malignancy. This does not mean it will not pass the borders in the future, it is not in that phase yet.

[2] Because the thesis covers several research fields the danger exists that incomprehensible jargon is used. For reasons of clarity and completeness I have added a thesaurus at the end of the thesis. All words which are explained in the thesaurus are written in bold face the first time they are used.

The breast can harbour two kinds of in situ malignant cells: ductal **Carcinoma** In Situ (DCIS) and Lobular Carcinoma In Situ (LCIS). The difference is in the place they are found, the first is found in the lobes, the latter in the ducts. For both there is no light-microscopic evidence of invasion through the basement membrane.

LCIS is the less dangerous one of the two kinds of in situ malignancies. If it is found no further therapy will be performed. LCIS is considered as the non-**comedo** type of DCIS, it rarely shows microcalcifications.

DCIS can again be divided in three categories of lesions:
1 the comedo type of DCIS has as discriminating feature the death of living tissue in the ducts. This process of dying of living tissue is called **necrosis**. Necrotic material becomes calcified. An important characteristic, which shows up on x-ray, films!
2 the cribriform type of DCIS with cells growing in seivelike pattern
3 the micropapillary type of DCIS which forms small tufts in the cavities of a duct and sometimes looks like a Roman aqueduct.

Both the cribriform and the micropapillary form grow in the lobules and should be called LCIS, but because the name LCIS was historically attached to the less dangerous form of malignancy. To show that these two forms of 'LCIS' have the same dangerous capacity as the comedo type DCIS, they were also called DCIS. A patient may exhibit one or more patterns of DCIS, sometimes in combination with LCIS.

3.Detection of breast cancer.

A very important component of cancer prevention and treatment is detection. The earlier a tumour can be found the easier the treatment. This is especially the case for malignant tumours. Early discovery can catch the carcinoma while they are still in their In Situ stage, where they have not spread through the body. After dispersion through the body there are several contaminated organs to be cured, only possible by an arduous and costly treatment as chemotherapy.

There are basically two ways which can lead to oncological breast examination The first has the ambiguous name BSE, short for breast self-examination. This should start as early as the age of 20 years. Woman should do this about once of month, also they should be examined by their doctor about once a year. Older women (40+) have a greater chance of getting breast cancer and should have a screening mammogram at regular intervals. This is the second way, which can lead to oncological breast research. Screening mammograms can show cancer before it can be felt and has a higher chance of detecting breast cancer. However this is a costly examination which makes it very difficult to perform regularly on all women past 20 years. Women in high-risk groups should be favoured

The risk groups include:
- Age, 75% of women with breast cancer is older then 50, under 30 it is very rare
- Family history, diagnosed cancer in mother, sister or daughter before menopause increases the risk.
- Personal history, women treated for cancer in one breast are likely to get cancer in the other
- Other risks, like early menstrual period, late menopause, getting children after 30 or getting no children at all.

The different test the doctor, or specialist can perform to examine the breast are:
- Aspiration: the extraction from a lump of matter by a doctor. Indicates whether it is fluid filled lump or a solid mass
- Mammography: examination of the breast by x-rays by a radiologist, x-ray film can show cancer before it can be felt but does not show all kinds of cancer (See Figure IV-2).

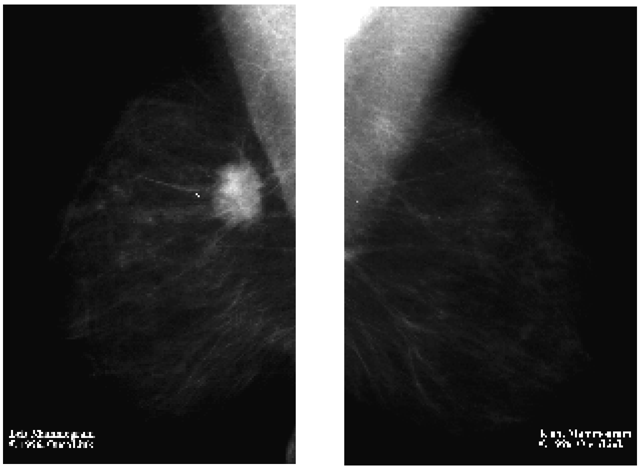

Figure IV-2: Mammogram clearly showing lesion in left breast.

- Ultrasound: Examination of the breast by high frequency sound waves. Pattern of echoes can show the difference between fluid-filled cysts and solid mass.
- Biopsy: Surgically removing part of or the whole lump, or suspicious area. A pathologist can examine the extracted tissue under the microscope. This is the only way to know for sure that there is a cancer. Biopsy can disfigure the breast.
- Hormone receptor test: if the biopsy shows cancerous tissue, tests can prove if female hormones stimulate its growth. If tested positive hormone therapy can be advised.

Biopsy is the best test to confirm that is a malignant tumour or for instance an innocent cyst, but surgery takes up the (valuable) time of the surgeon and can be damaging to the appearance of the breast causing psychological anguish to the patient (See Figure IV-3). The best option would be a regular mammogram screening for all women past a certain age, preferably 20. The low radiation dosage is not harmful for the breast, carcinoma may be found in an early stage and biopsy is only necessary after the mammogram confirmed the malignancy of the lesion. This again takes up the time of the radiologist making it a very expensive program. Automating the process to make computer classify x-ray films would be time- and cost saving. The following section elaborates on how mammography and diagnosis of the films is performed.

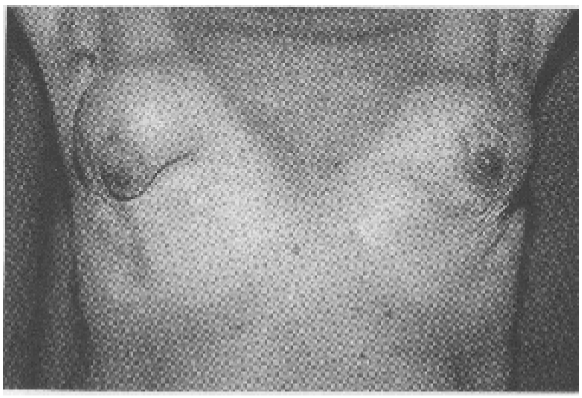

Figure IV-3: Woman, aged 42, has had 3 biopsies because of microcalcifications. Both breasts are disfigured and woman reports sexual problems and rejection by her partner.

4.Mammography and microcalcifications.

Making x-ray films of the breast is called mammography. Each breast will be imaged in two directions: from the top facing down and from the side. All pictures together compose the mammogram. The breast is hereby squeezed by the special mammography machine to provide the clearest picture. This machine produces a low dose x-ray, which passes through the breast and forms the image. The image has to be taken be a special machine because the different tissues of the breast have very similar densities, distinguishing among them requires a special x-ray beam and film.

Figure IV-4: Making a mammogram.

A difference has to be made by normal mammography and screening mammography. Normal mammography takes place if a lump is already detected in the breast or there is a high chance there could be a carcinoma (e.g. one was found in the opposite breast). In this case mammography is used to confirm a possible diagnosis. Screening mammography only detects presumptive abnormalities. After these have been detected the patient undergoes further examination where additional techniques may be applied (biopsy). The research in this thesis is primarily focussed on screening mammography.

Screening mammography requires advanced diagnosis techniques because after all the diagnosing radiologist has to detect possible carcinomata in a very early stage. This means that in most cases there will be no clear lesion to show up at the X-ray film. He has to look at the details and gather the small pieces of information to make a diagnosis. The problem is that X-ray film only shows contrasts in the tissue. For something to distinguish itself it has to be opaque to X-rays. Also the X-ray film has to have a very high resolution to show up all the details. The microstructures to base the diagnosis on are extremely small fibrotic structures and microcalcifications.

Microcalcifications

In 1913 Salomom [1913] took X-ray films from amputated breast and noticed that there were 'small black spots' in the centre of a carcinoma. He diagnosed it as 'intraductal cancerous masses undergoing cystic degeneration'. These spots were the first microcalcifications ever to be detected radiographically. Microcalcifications are small calcium deposits.

The origin of these intramammary calcifications differs. They can be the result of inflammatory, degenerative and toxic metabolic processes or mechanical injury. A hypothesis for the pathogenesis of the calcifications by carcinoma is proposed (among others) by Stegner and Pape [1972]. Theory of Stegner and Pape is that the permeability of the cell membrane to calcium ions is greatly increased in actively secreting ductal carcinomata, so that the calcium deposit can grow in a optimal pH environment.

The pluriformity in origin shows that carcinomata are not the only ones producing microcalcifications. Detection of microcalcifications may indicate the dangerous DCIS but also harmless milk of calcium cysts. If the radiologist cannot tell the difference it will have to be determined by a biopsy. At the Cologne University Women's [Citoler 1978] Clinic, 90 percent of performed operations after detection of microcalcifications was unnecessary. Lanyi presents in his book a differential diagnosis system, which improves these figures to 27 percent; i.e. 73 % of women

with microcalcifications were exempted from biopsy [Lanyi 1986]. Of 42 diagnosed carcinomata only one was diagnosed as benign (for a complete checklist see appendix A).
Perfect diagnosis of microcalcifications is impossible but there are some general guidelines [LeGal et al 1976, 1984]:

1. The more irregular, linear, or wormlike the calcifications, the higher the risk.
2. Microcalcifications arranged in linear or branched pattern signify a high risk.
3. Rounded or ringlike calcifications with evidence of sedimentation (teacup) are not suggestive of malignancy (See Figure IV-5).

Figure IV-5: microcalcifications with teacup sign, diagnosis harmless.

Microcalcifications originate for 75 % from benign lesions and 25 % malignant. The majority of malignant lesions emanate from Ductal Carcinoma In Situ (DCIS). Differentiating between malignant and benign is done by analysing their form, size, number and distribution.

Taxonomy of origin of calcifications [Frieling 1992]:

1. Lobular
   a) cystic breast disease (mastopathy)
   Adenosis is the excessive growth of (gland) tissues in the lobules. If the lumps contain small amounts of calcium the condition is called sclerosing adenosis.
   - a micro-cystic (blunt duct) adenosis
   - b sclerosing adenosis
   - c micro-cystic adenosis

   b) *lobular neoplasia* (LCIS)[3] [4]
   c) *infiltrative lobular carcinoma*

2. Ductal
   a) ductal carcinoma in situ (DCIS)
   - *a comedo type*
   - *b micropapillary type*
   - *c cribriform type*

[3] malignant lesions are in italic
[4] rarely show microcalcifications

b) retention of secretion with or without plasma-cell mastitis (Secretory disease or benign duct ectasia).
c) intraductal fibroadenoma (See fibroadenoma)

3. Fibroadenoma.
   Fibroadenoma are estrogen induced benign tumours. They are made from fibre and gland tissues (aden means gland in Greek). They feel rubbery, move around easily. Fibroadenoma are the most common type of tumours in women in their late teens and early twenties.

4. Other
   a) fat necrosis (lumps formed by damaged and disintegrating fatty tissue).
   b) arteries
   c) parasitic disease
   d) foreign bodies
   e) sebaceous glands
   f) lymph nodes

CAD or CAR

As in most fields of science computers have entered the field of mammography. It is called Computer Aided Diagnosis (CAD) or Computer Aided Radiology (CAR). The simplest form of CAD being the digitisation of mammograms which allows the radiologist to zoom in, enhance, transpose the images to allow a better view. An interesting development is the digital stereomammography were the radiologist is given a 3-dimensional view by stereoscopy [Chelberg 1994].

A more radical approach was chosen by Nico Karsemeijer [1993]. He developed a statistical method where the computer rescales the images to equalise image noise. The rescaled images had much better results then the original ones. Th computer uses a Bayesian algorithm to recognise and segment the microcalcifications. The microcalcifications or clusters of microcalcifications are then zoomed in on to simplify the classification. An interpolation method is used to compensate for too low graphic resolution of the images. This way a radiologist can concentrate on the specific area were the microcalcifications are found.

The microcalcifications are diagnosed mostly diagnosed in clusters, but the question arose if maybe individual calcifications could be a clue to possible diagnoses. After all microcalcifications are often classified by their morphology. W. Veldkamp tried to classify individual microcalcifications with a statistical method (k-nearest neighbour). Karsemeijer and Veldkamp reckoned that maybe neural networks could do the job. They have several sets of classified mammograms on which they used their segmentation method to isolate the microcalcifications. So there is a set with inputs paired with their desired outputs, which makes it an ideal candidate for artificial neural classification. This is the main goal of this thesis. Research if neural networks (in this case backpropagation) can classify individual microcalcifications on malignity.

If computer classification would be possible then it would be a useful tool for radiologists. Normally a mammogram is examined by several radiologists to minimise the chance on human error. Maybe if the artificial classification worked within a certain error margin then it the number of radiologist could be reduced to save time. The network could be used as an extra check after the diagnosis has been determined.

5.Treatment

When cancer is diagnosed there a number of possible treatments. To choose a treatment there are a number of parameters to consider. The stage of development of the tumour, the type of tumour, the age of the patient, the menstrual cycle and the physical condition of patient can all influence the treatment decision. If the circumstances allow it then the patient itself may choose from a number of options the team of physicians offer. She[5] can choose which treatment or combination of treatments to choose by weighing the risks involved by each treatment.

The treatments roughly divide into four categories: surgery, radiation therapy, chemotherapy and hormone therapy. Keep in mind that combinations of different categories may be used to treat a patient, e.g. the surgical removal of a tumour and succeeding chemotherapy to remove possible cancer cells which have spread through the rest of the body. Here they are in more detail:

Surgery

Surgery is most frequently used treatment of breast cancer. The surgeon operates on the patient to remove the cancer from the breast and usually some lymph nodes under the arm. There are several operations possible:

Lumpectomy: operation (also called partial mastectomy, segmental mastectomy) in which only the lump is removed from the breast and the lymphnodes are removed under the arm. Also little tissue around the cancer is removed from the breast. Usually additional radiation therapy and sometimes chemotherapy or hormone therapy is needed.

👍 breast does not have to be removed.
👎 additional x-ray therapy is needed.
👎 possible arm swelling.

Mastectomy: operation (also called total mastectomy) in which the entire breast is removed along with some skin. If also the underarm lymph nodes are removed it is called 'total radical mastectomy'.

👍 usually no additional therapy is needed.
👎 breast and surrounding skin is removed causing severe aesthetic disfigurement.
👎 possible arm swelling.

Radiation therapy

A radiation oncologist uses high doses of radiation to kill any cancer cells. Radiation may come from on outside source or a small tube may be implanted in the breast to provide the radiation.

👍 breast does not have to be removed.
👎 long treatment (5 days a week for 5 to 8 weeks).
👎 there may be long term changes in the appearance and texture of the breast and skin of the breast.
👎 possible arm swelling.

Chemotherapy

Oncologist gives drugs, which kill the cancer cells anywhere in the body. Numerous drugs are available and often a cocktail of drugs is used.

👍 treatment can kill any unsuspected cancer cells spread around the body which can not be treated by radiation therapy or surgery.
👎 unpleasant side effects as temporary loss of hair, loss of appetite, vomiting and tiredness. When the ingestion of the drugs has ended, the side effects subside.
👎 possible temporary or permanent sterility.
👎 harmful for foetus so unusable during pregnancy.

[5] Breast cancer does occur among the male population but this is so rare and detection techniques are so different that it falls out of the scope of this chapter

Hormone therapy

The progression of certain cancer types can be stimulated by hormones produced by the body. The doctor administers drugs which change the way these hormones work and stop the growing of the cancer cells. As in chemotherapy cancer cells can be stopped growing anywhere in the body.

- 👍 therapy drugs are in the form of pills.
- 👎 can cause infertility, hot flashes, vaginal dryness and other symptoms of menopause, in spite of age.

6.Conclusions

Breast cancer is one of the most common fatal diseases for woman today. Several treatments are developed to reverse the effects of malignant tumours but are costly (surgery, x-rays) and can have burdensome effects on the patient (chemotherapy, mastectomy). Detecting the carcinoma in an early stage is important to prevent the disease form spreading through the body. This detection is best done by mammographic screening of the female population, especially women falling in high-risk categories. This large screening generates huge amount of x-ray films all to be diagnosed by radiologists.

Mammographic screening is a preventive method, so possible malignancies have to be detected in a very early stage. Microcalcifications can indicate the development of a malignant tumour and are visible to X-ray, but microcalcifications can be created by other (less harmful) diseases as well. There are methods of classifying these microcalcifications but they still have high error margin (50 %). This means for those incorrectly diagnosed malignant, needless stress for patient and family and a possibly disfiguring biopsy.

Computer Aided Diagnosis can assist to enhance the mammograms. This can attribute more information for a certain diagnosis. A special method has been developed which isolates and magnifies microcalcifications to simplify classification. The subject of this thesis is the classification of those isolated microcalcifications by a neural network and will be treated in the next chapter.

Chapter V: Research

Mammographic screening produces enormous amounts of mammograms. All these mammograms have to be examined by a radiologist and checked by another. This process is time consuming and expensive (radiologist gets paid by the number of patients). Automating parts of the process would speed up examination. The research performed for this thesis investigates if neural networks can assist in automating the process.

There are now several applications that help the radiologist make his diagnosis. Most of them use digital image processing. The mammograms are digitised so the radiologist can look at them on his monitor. He can zoom in, turn, and change contrast and so on. Further going is Karsemeijer's segmentation program, which isolates and magnifies areas where microcalcifications are found, so the radiologist can concentrate on the microcalcifications only.

All these applications are very passive. They give the radiologist full control over the examination and diagnosis process and perhaps this should stay this way. Radiologist are in general not too happy allowing computers to do part of their job and most people do not trust computers to make decisions which can so evidently change their lives. But screening mammography is as yet not without error. Looking at microcalcifications 50 percent of all cases diagnosed as malignant are in fact benign lesions. This means a large number of women who have to undergo an unnecessary biopsy to rectify the diagnosis. So maybe more active applications are needed.

One of the possible active applications involves the classification of the microcalcifications. After the segmentation algorithm detects them they can be checked on distinguishing features by a computer. Although radiologist look mostly at clusters of calcifications, maybe information can be found be examining individual ones. This thesis studies if the classification can be done with the help of neural networks.

Maybe the network can find features hidden in the pixel values of the small images of the microcalcifications. This would be a great help to radiologist in actively assisting in the diagnosis. It could be used to double check a diagnosis or filter the huge amount of mammograms by removing the clearly harmless ones. The question if such an active application should be used, will I leave others

Summary of sections

I will start of with section with information about the dataset. Where is it from, what is the layout of the set and how was it made? Next will be the visualisation of the microcalcifications. Maybe there is a pattern to be seen with the naked eye. This pattern could then be emphasised by pre-processing the dataset before it is fed to the neural network. This pre-processing of the dataset and making it digestible for the backpropagation network will be dealt with in the next section, followed by a description of backpropagation implementation and the parallel computer system it runs on.

The core part of the research will be explained in the second half of the chapter. First the network will be trained and if the results are satisfying it will be tested. The chapter will conclude with the analysis of the results.

1.The datasets.

The first dataset I received contained the information about 72 cases. The set originates from the archive of the LCRB (Landelijk Referentie Centrum Borstkankerscreening, The Netherlands). Presumably they are screening mammograms, i.e. from women who participated in the screening program. All cases were diagnosed by a pathologist and classified into two categories, malignant and benign. Be aware that the classification is based on the whole case, i.e. the pathology of the breast itself. Each case contains a number of images of microcalcifications from the breast. Each microcalcification is 2.5x2.5 cm^2. The selected images were digitised at a resolution of 0.1 mm per pixel. An expert radiologist would mark out the main area where the microcalcifications were seen and the segmentation program [Karsemeijer 1993] would isolate them and add them to the set. This means all the microcalcifications found on one mammogram have the same classification, which could be a problem for the supervised backpropagation. Backpropagation needs a correctly classified dataset for learning.

Later on it became known that the dataset was build up from mammograms taken over a period over ten years. In these years the quality and resolution of the mammography machines improved, this changed the contrast of the images. The change contrast showed up in the mammograms and unbalanced the dataset. Even more disturbing was the fact that all the benign cases were taken recently, this way the network would just have to look at the contrast to classify the microcalcifications. Veldkamp devised a logarithmic adjustment method to cope with the problem but in the end it proved unnecessary because a new, better-balanced, case-database became available. This dataset contained only 56 cases but all were taken with the same contrast.

De dataset contains the following information:
Per set:
- The number of microcalcifications

Per microcalcifications:
- 15 by 15 greyscale image (pixel values between 0 and 4095).
- 15 by 15 black and white demarcated image (pixel values 0 and 1) [6].
- Current calcification number.
- Code if the calcification is bigger then the 15 by 15 field.
- Type, the classification provided by the pathologist.
  0 = benign
  1 = malignant (comedo type)
  2 = malignant (non comedo type)
- Y-co-ordinate of calcification on mammogram.
- X-co-ordinate of calcification on mammogram [7].
- Filename of corresponding mammogram.

[6] The differentiation of the microcalcifications from the background tissue is done by the segmentation program and can a bit arbitrary. It depends on the threshold parameter of the segmentation program. The value of this threshold is inversely proportional to the number and size of calcifications

[7] The co-ordinates of the microcalcifications can provide information on the possible diagnosis, for instance one could calculate if the calcification is pointed to the nipple, which can be a clue for malignancy. This information is not used in the neural network classification because it would also need the co-ordinates of the nipple itself, which were not provided. It would also be difficult to make the network notice the distinction between the pixel values and the co-ordinates.

2.Visualising the calcifications.

Working with rows and rows of pixel values is difficult and complicates understanding of the

Figure V-1: Greyscale image real size

problem. By visualising the images you can see for yourself what kind of pattern can be detected and if there is a pattern to be detected. For this a small program had to be created which could extract images from the case-file and enhance the images created.

The program created reads the pixel values from the case-file and converts them to a picture file (PGM format). It has parameters, which blows up the image to an acceptable size and a parameter what kind of image to produce, the greyscale picture or the black and white demarcated image.
Both pictures clearly show a microcalcification in the middle with another at the edge of the image.

Figure V-2: Greyscale image enlarged 15 times.

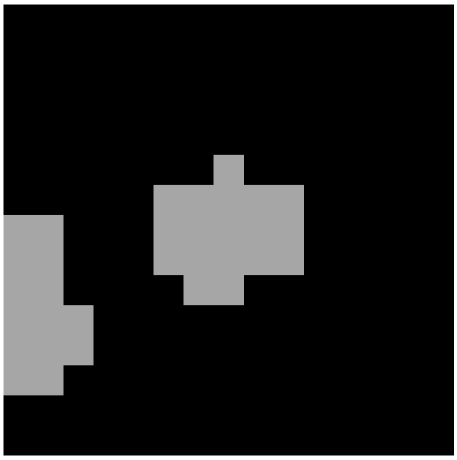

Figure V-3: Binary image enlarged 15 times

They are a clear example how small these microcalcifications are, who low the graphics resolution of the ultimate image is. To improve the image I tried two interpolation methods, (See Figure V-4 [Keys 1981]) to produce a sharper picture but results were unsatisfactory. Both methods performed bad around the edges and delivered pictures, which were nice to look at but hardly an interpolation of the real image.

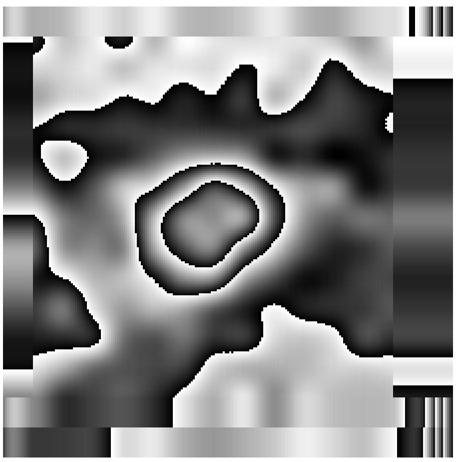

Figure V-4: Greyscale image enlarged with cubic interpolation.

It is hard to recognise and classify the microcalcifications from an image. The graphic resolution of the digitised mammogram is very low. The images were digitised using a 12-bits CCD-camera. This resolution is okay for the whole picture, but looking at the picture in detail (like a microcalcification that generally has a size of about 2 by 2 pixels) gives a poor quality.

3.Converting the dataset

Data-wise the case-database has all the data needed to train and test the backpropagation network. It has a large number of input vectors all paired with a desired output (1,2 = malignant, 0 = benign). To convert the microcalcifications-data to a test-set and a train-set again a small program was made a modification of a program made by Schouten.

The program starts by reading all the vectors (images) and classifications from the data-files. The values of the pixels are scaled to the domain of the output function. The domain [0:4095] is scaled to the domain [-1:1], this is taken instead of [0:1] because it usually yields better results. All vectors are divided by category and then in turn divided over the test-set and train-set. To give more control a number of parameters are added to pre-process the dataset.

As mentioned earlier in the first two chapters, on the learning domain neural networks are very difficult to control and to monitor the progress of what they learn. The only monitoring tool is the error values but they only give information that the network is learning something. That 'something' does not necessarily have to be the desired I/O-function. To give more control to the user, a number of parameters are added to the conversion program to pre-process the dataset:

- **type**: the type of image distinguishes between the values of the greyscale images and the values of the binary image. The binary image only shows the form of the microcalcifications, forcing the network to ignore subtle differences in contrast.

- **window**: as mentioned the calcifications are 15 by 15 pixels big, with this parameter you can restrict this size to smaller values, by placing a window over the original values. This reduces the size of the first layer of the network (from 255 to 81 for a 9 by 9 field). This reduction of input neurons decreases the computation thereby speeding up the network. It also it masks possibly nonessential background pixels (the microcalcifications is always located in the middle of the image).

- **removal**: some microcalcifications are so close to others that part of the other microcalcifications appear in the image. Applying this parameter removes this other microcalcifications. For the binary image this just means masking all pixels of the extra calcifications. For the greyscale image the pixels are replaced by random values which are between the highest and lowest background values.

- **size**: this parameter applies only for the binary images. Because image resolution is rather low, the case-files contain a number of microcalcifications with only one or few pixels marked. This creates a number of identical binary input vectors with different classifications. Obviously these identical vectors severely complicate network learning. Choosing a value causes all microcalcifications smaller than that value to be ignored. NB this reduces the number of input vectors, which makes it more difficult for the network to learn the I/O-function.

4.Backpropagation and the PowerXplorer

Artificial neural networks are modelled on the biological prototype. This prototype, the human nervous system, has unparalleled computation power by virtue if one main feature, massive parallelism. Consequently artificial neural networks are not difficult to implement on a parallel computer. The computation done by the neurons can be divided over a number of processors to speed up network training.

The architecture chosen for this implementation of backpropagation was the PowerXplorer by Parsytec [8]. It is a system of 16 parallel processors grouped in partitions of four with a Unix shell to control it. The system is monitored, controlled and connected to the network by a Sun workstation. The number of processors to use is optional. This means that when you choose only four processors (1 partition), other users can utilise the remaining processors. Minimum usage is one partition, so there is a maximum of 4 users at the same time. For more specifications see appendix B)

The backpropagation implementation on the PowerXplorer was written by Theo Schouten. It has an interactive as well as a batch mode. The interactive mode can be used to fine tune the network and examine the results (e.g. adjusting the step size by changing the learning rate and see if it has any effect on the error value). The batch mode can be used to run large batches of network training. This is convenient if the network training time takes a very long time (can take days or weeks) or you want to try out the network with different parameters.

The program contains numerous parameters, which can be adjusted runtime (interactive mode) or design time (batch mode). Here are the most important ones:

- **nlayers**: number of layers for the network (3 layers means, 1 input layer, 1 hidden layer and 1 output layer).
- **neurons**: number of neurons in the network (225 100 3 means 225 input neurons, 100 hidden neurons and 3 output neurons), I modified the program to allow the user to change the number of neurons in the first hidden layer in the batch-language.
- **netmode**: chooses between different network modes. First there are two ways of learning, on-line and epoch learning. On-line you calculate the delta and immediately change the weight, while epoch learning collects all deltas and changes the weights after an epoch. The usage of parallelism has also two forms, distributed net and data set parallelism. Distributed net means you divide the neurons over the processors and each processor performs the calculation for its own neuron. Data set parallelism means the dataset is divided over the processors. Each processor gets all the neurons but computes only its own part of the dataset (weight changes are collected, combined and distributed).
  ONLINE_NET, distributed net, on-line learning.
  DIST_NET, distributed net, epoch learning.
  DATA_NET, data set parallelism (only works for epoch learning).
- **learn_method**: chooses between different backpropagation learning methods:
  0 learning without momentum
  1 learning with momentum (method used)
  2 quickprop, an optimisation in speed for backprop
- **learn_eta**: η, learning rate, step size (must be between 0 and 1)
- **learn_momentum**: α, momentum, involves the previous change to calculate the current changes (must be between 0 and 1).
- **maxiter**: the maximum number of iterations the network learns before it stops (when it does not reach the desired error-value).
- **randomfactor**: effect the precision of the first random initialisation of the weights
- **activation function** (only runtime), choose the way the NET-value is squashed. All have parameter s, which effects the steepness of the function (can also be changed). Possible parameters are:
  sigmoid, 1.0 / (1.0 + exp(- s u) (domain [0:1]).
  hyperbolic tangent, tanh(s u) (domain [-1:1]).
  Elliot's function, (s u) / (1 + abs(s u)), fast approximation of the tanh[9].

[8] For more information on the PowerXplorer and Parsytec bring a visit to their homepage on the Internet http://www.parsytec.de.

[9] because the sigmoid and the hyperbolic tangent are implemented by table lookup, the speed advantage of Elliot's is very small (a few percent at most). It will not be used in this thesis

- **confusion_delta**: maximum difference between output value and closest target value of a certain class to confine an output to a certain class. If the difference between the output value and target value is higher then the delta then the output is classified undetermined[10].

5.Training

Tuning the network.

A major problem with neural networks is finding the right parameters. For instance how many neurons does it need in the hidden layer, should I use momentum and what value for momentum, how big is the step size? The optimal values for the parameters can vary with the dataset and although there are some general guidelines (e.g. start with high learning rate and slowly decrease it); one has to tune the network each time there is a new dataset.

To try to automate this process of tuning the network I created a program, which would permute a number of values for the main parameters. The program generates a batch-file that would run the network for all permuted configurations and save the results. This way one can find the best parameters by comparing the error values. I choose three main parameters: learning rate, momentum and number of neurons in the first (only) hidden layer.

These are the results for the greyscale 15 x 15 images in the dataset without any pre-processing. The network has 255 input neurons and 3 output neurons. The network runs 1000 iterations and uses data set parallelism. The learning dataset contains 1311 patterns (as does the test set).

Tabel V-1: The parameters for the permutation program

neurons	10	100	200
eta	0.1	0.05	0.01
momentum	0.8	0.1	0.0

Table V-2: Results of dataset with greayscale images

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
10	0	0.1	986	0.494	1.640	0.939	159	5.645	1.297	93.059	1.0	93.0
10	0	0.05	975	0.505	1.588	0.951	180	3.280	1.220	95.500	1.0	95.4
10	0	0.01	954	0.519	1.572	0.969	668	1.678	1.526	96.796	1.2	96.7
10	0.1	0.1	972	0.497	1.640	0.946	180	4.195	0.992	94.813	0.8	94.8
10	0.1	0.05	951	0.504	1.588	0.955	180	3.204	1.220	95.576	1.0	95.5
10	0.1	0.01	972	0.517	1.575	0.968	668	2.136	1.449	96.415	1.2	96.3
10	0.8	0.1	71	0.518	1.657	0.967	159	2.288	1.297	96.415	1.0	96.4
10	0.8	0.05	988	0.467	1.663	0.897	159	11.823	1.983	86.194	1.3	86.6
10	0.8	0.01	981	0.492	1.640	0.942	180	4.424	0.839	94.737	0.6	94.7
100	0	0.1	963	0.494	1.562	0.937	1237	4.653	0.610	94.737	0.5	94.7
100	0	0.05	907	0.499	1.613	0.947	1237	3.432	0.534	96.034	0.4	96.0
100	0	0.01	974	0.514	1.599	0.964	668	1.907	1.144	96.949	0.9	96.9
100	0.1	0.1	945	0.493	1.541	0.936	1237	4.882	0.915	94.203	0.7	94.1
100	0.1	0.05	945	0.498	1.618	0.953	1237	3.661	0.534	95.805	0.4	95.7
100	0.1	0.01	927	0.513	1.618	0.958	668	2.288	1.297	96.415	1.0	96.4
100	0.8	0.1	965	0.464	1.779	0.885	159	12.891	1.449	85.660	0.9	86.3
100	0.8	0.05	970	0.469	1.561	0.884	159	11.365	1.602	87.033	1.2	87.6
100	0.8	0.01	957	0.493	1.623	0.947	1237	4.348	0.686	94.966	0.5	94.9

[10] Except for the testing results this parameter is always set to 0.5, which means that is the closest target value has a difference of more then 0.5 then it will be classified as undetermined.

200	0	0.1	991	0.495	1.557	0.942	1237	4.577	0.839	94.584	0.6	94.5
200	0	0.05	939	0.501	1.563	0.945	668	3.432	0.763	95.805	0.6	95.8
200	0	0.01	976	0.511	1.610	0.962	668	1.907	0.992	97.101	0.8	97.0
200	0.1	0.1	973	0.494	1.567	0.937	1237	5.034	0.763	94.203	0.5	94.1
200	0.1	0.05	916	0.499	1.579	0.948	668	3.738	0.686	95.576	0.5	95.5
200	0.1	0.01	958	0.510	1.615	0.957	668	2.059	0.686	97.254	0.6	97.2
200	0.8	0.1	997	0.442	1.732	0.845	725	17.849	2.365	79.786	1.5	81.2
200	0.8	0.05	984	0.457	1.657	0.886	159	11.747	2.136	86.117	1.5	86.8
200	0.8	0.01	946	0.485	1.598	0.922	1237	5.873	0.915	93.211	0.6	93.3

The time to calculate these data was 31 hours and 34 minutes, which averages to ca. 70 minutes per configuration. Keep in mind that increasing the number of neurons as increases number of weights. Doubling the neurons doubles the training time. So the nets with 10 neurons took less time and the nets with 200 neurons took more time then 70 minutes.

Explanation of column names:

- **Neu** : number of neurons in the hidden layer
- **Eta** : learning rate
- **Mom** : momentum
- **Iter**: number of iterations performed since start after reaching lowest error
- **err** : current error
- **maxerr**: maximum error on any output neuron for all patterns
- **pat**: pattern number where this occurs
- **averr** : average of maximum error per pattern
- **%ok,%bad,%und**: consecutively percentage of correctly, wrongly and not classified patterns (is influenced by confusion_delta parameter).
- **rmsbad,rmsund**: consecutively percentage of bad, undecided root-mean-square entries in confusion matrix

Table V-3: Top 3

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0.8	0.1	997	0.442	1.732	0.845	725	17.849	2.365	79.786	1.5	81.2
200	0.8	0.05	984	0.457	1.657	0.886	159	11.747	2.136	86.117	1.5	86.8
100	0.8	0.1	965	0.464	1.779	0.885	159	12.891	1.449	85.660	0.9	86.3

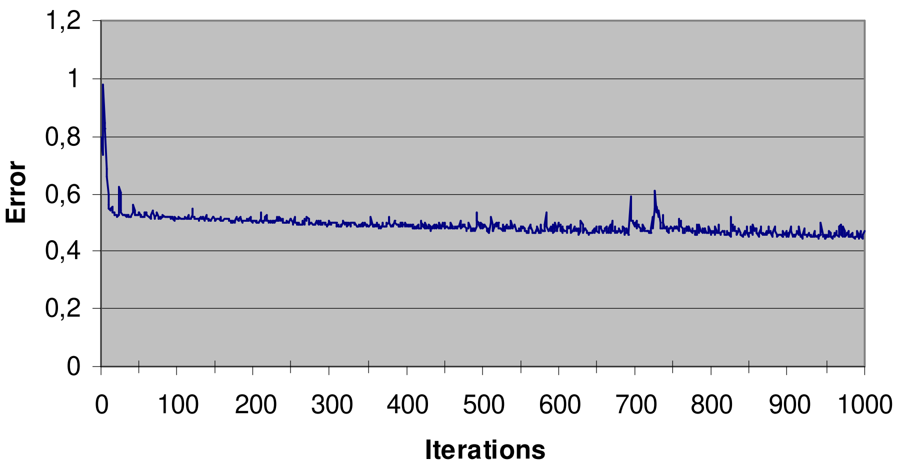

Figure V-5: Error value of best result

From these results one can conclude that 200 neurons in the hidden layer are a satisfactory amount (10 gives much worse results and 100 gives a slightly worse result, NB the number of neurons should be kept as low as possible to restrain training time). Furthermore a high value for momentum (in this case 0.5) is preferred because no or low momentum slows down network learning. For the learning rate the best results vary, this is because the high value, which scores the best in this run, does not have to be the best value. High values for learning rate can speed up network convergence in the beginning but stop learning when it reaches a certain precision. Then the learning rate has to be lowered to further minimise the error. Choosing a low learning rate at the start will only slow down convergence. Figure V-5 shows the error value of the best result (neurons 200, momentum 0.8, eta 0.1). Although the learning rate already reached a low value, it is still descending after 1000 iterations.

6.Pre-processing the dataset.

De dataset contains a huge amount of data. Probably not all the information that is in the dataset needed, some information may even be confusing to the network. Also a large dataset means a long training time. So several new datasets were tested if they could produce better results.

Binary pictures.

Instead of the greyscale images the dataset is made from the binary images. The input neuron will receive a -1 if the pixel belongs to the background, the neuron will receive a 1 if the pixel belongs to the microcalcification. All the information the network gets is a black and white bitmap of the microcalcifications. The idea behind this conversion is that the main feature the radiologist looks at is the shape of calcification. By feeding the network the binary image it is accomplished that all the information the network gets is the morphology of the microcalcification, features like for instance contrast are eliminated.

Dangerous for this type of classification is that the images are already on the maximal resolution. Noise or an approximation error can cause differences between neighbouring pixels. A small difference can just be enough to pass the threshold of the segmentation function changing the pixel from background to microcalcifications.

The below best results were produced by the same batch as above results from the normal dataset. That means the same parameters and 1000 iterations.

Table V-4: Top 3 binary pictures

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
100	0	0.01	84	0.513	1.628	0.975	93	8.162	0.839	90.999	0.8	92.3
100	0.1	0.01	68	0.522	1.583	0.982	93	6.255	1.297	92.449	1.3	93.6
10	0.1	0.01	465	0.543	1.566	1.012	1234	3.738	1.297	94.966	1.1	95.5

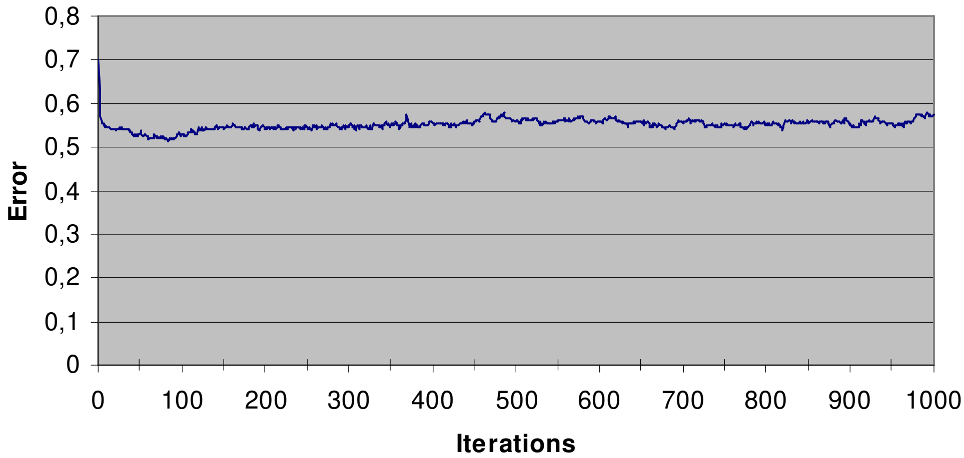

Figure V-6: Learning rate of binary pictures (best result)

Evaluation of the results clearly shows failure to learn. The best values are found with using low or no momentum and high learning rates. Also the error-curve fluctuates heavily showing a malfunction in the convergation process. Best value is reached after 84 iterations while 1000 iterations where performed (See Figure V-6). Closer examination of the dataset reveals numerous images from different categories with only one pixel marked as microcalcifications and the rest marked as background. Because the middle of the microcalcifications is centred in the middle of the image there are many identical input vectors with different classifications. In short the dataset is corrupted and impossible to learn.

Size

Closer examination of the binary images showed several images from different classes with only one white pixel. These microcalcifications are simply to small to extract information from. To solve this problem an extra pre-processing method was conceived. Remove all the calcifications from the dataset whose binary images contains less then x pixels. Only microcalcifications with a certain size remain with a clearer morphology and less lightly to contain grave corruption by noise. Disadvantage of this method is that it reduces the dataset [see table], increasing minimal size from 1 to 4 halves the number of input vectors. While this does decrease training time for one iteration, it frustrates the training process by removing information. The more data to extract information from the better.

Table V-5: Size versus number of input vectors.

size	training set	test set
1	1311	1311
4	775	774
10	318	316

Table V-6: Top 3 for binary images minimal size 4.

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
100	0.1	0.01	100	0.452	1.636	0.875	764	17.419	2.065	80.516	2.3	83.4
200	0	0.01	40	0.461	1.515	0.881	584	17.806	2.839	79.355	3.1	84.0
100	0	0.01	92	0.466	1.596	0.896	432	16.516	1.419	82.065	1.7	85.1

Strangely the network reaches relatively low values after few iterations with zero or no momentum. Somehow this net learns better without momentum. To maximise this method the size was increased and an extra modification was performed. This produced satisfying results that could be used for further learning and testing (see next section).

Removing extra calcifications.

As shown in Figure V-2 calcifications can lie so close to each other that they show up on each other's images. Especially for binary images this is an unwanted side effect because the network is trying to recognise certain features in the morphology of the microcalcifications. But having another microcalcifications in the neighbourhood can also be an important feature so caution is important. This should be possible for the greyscale image as well as the binary images. For binary images the operation is simple: just change all pixels from microcalcifications, which are not in the centre to zero.

For the greyscale images it is more complicated because the pixels from the removed microcalcifications have to be replaced by pixels, which fit into the background. A method that replaced the pixels with random background values showed heavily deteriorated learning results. Apparently the method left too much indications of the alteration, which disturbed the training process. This is a difficult problem because the best way to replace the pixels would probably be some kind of interpolation, but if this interpolation would work correctly it would just put the microcalcifications back into the image.

Table V-7: Top 3 for binary images with minimum size 10 removing extra microcalcifications.

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0	0.01	77	0.415	1.519	0.795	298	27.987	5.031	66.981	6.4	73.9
100	0	0.01	186	0.442	1.512	0.844	51	21.069	2.516	76.415	4.0	82.1
100	0.1	0.01	154	0.448	1.564	0.847	32	22.013	4.403	73.585	6.1	78.1

Remarkable is the very fast learning time for bigger network, which could point to a total memorisation of the network. This means the network does not recognise features but just stores the whole image in the network. The bigger the network the better the storage capacity. Also the reduced dataset makes it easier to store it into memory. The very low error value makes it an excellent candidate for further testing.

Reducing size of image

Most microcalcifications are very small, even the ones consisting out of more then 10 pixels are still relatively small compared to the background. A very big part of the input vector will be part of the background. That is strange because it were the microcalcifications we are interested in. A way to emphasise the microcalcifications more is to 'cut off' parts of the edges of the image. Remember that the calcification is in the centre of the image so it will not be discarded. Most microcalcifications will still fit, except for a few microcalcifications that actually overlap the edges of the image. A very handy side effect is the reduction of the number of input neurons. In this case a 9x9 field will need only 81 input neurons instead of 255 thereby severely cutting down on the training-time (33 % of the original time).

The parameters for the permutation program:

neurons:	200[11]	0.1	0.01
momentum:	0.8	0.0	

[11] The largest network seems to work the best and reducing the size for time saving in the training process is not necessary because low training time.

Table V-8: Normal with window 9x9

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0	0.01	500	0.516	1.761	0.949	10	9.077	2.670	88.253	1.6	88.5
200	0.8	0.01	6	0.546	1.671	0.988	158	0.305	0.610	99.085	0.4	99.1
200	0.8	0.1	166	1.369	1.800	1.423	1150	32.113	50.801	17.086	26.6	17.6
200	0	0.1	128	1.375	1.800	1.361	323	31.579	55.835	12.586	30.1	12.4

Table V-9: Binary with window 9x9 size 4

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0	0.01	322	0.315	1.648	0.653	169	44.000	1.806	54.194	1.6	59.1
200	0.8	0.01	2	0.551	1.501	0.997	654	2.581	1.161	96.258	1.3	96.8
200	0	0.1	124	1.200	1.800	1.411	260	25.677	44.903	29.419	25.7	29.3
200	0.8	0.1	370	1.260	1.800	1.403	659	33.806	44.129	22.065	23.7	22.6

Table V-10: Binary with window 9x9 size 4 with removal

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0	0.01	475	0.335	1.569	0.684	769	40.387	2.452	57.161	1.5	61.7
200	0.8	0.01	2	0.550	1.500	0.997	157	3.097	1.032	95.871	1.1	96.6
200	0	0.1	13	1.212	1.800	1.336	445	33.290	46.710	20.000	24.4	19.9
200	0.8	0.1	370	1.305	1.800	1.433	491	32.774	44.387	22.839	24.4	23.5

Strangely all configurations reach their lowest error value with no momentum. All training runs with high momentum immediately get worse (the combination momentum 0.8 and learning rate 0.01 start with a low result which immediately goes up). Classifying the binary images produces better results

Adjusting learning rate.

Rumelhart, Hinton and Williams [1986] proved that backpropagation will always converge, as long as you use infinitely small steps. Only problem was that it would take infinite training time. A more practical conclusion is that very low values for learning rate makes sure the network will keep learning and decrease the error value. Problem is it takes to long.

Training the network is a bit like driving a car to a friend. When you are a long way from home you use the highway with great speed and high gear. When you arrive at his town, you use the town roads with average speed and middle gear. When you reach his street and want to park you use low speed and gear. Converting the analogy to networks would be changing gear to learning rate and driving speed to learning rate. When you're a long way from you use high learning rate and the closer you get the lower the learning rate and learning rate.

Easiest way to do this is by continuously watching the network in interactive. As soon as the learning error stops descending (or even starts ascending) lower the learning rate and wait for the next stop. This does take time and can mean staring at a computer screen for a long time. Maybe in the future a system like gearing system in the old Dutch DAF-cars (and in most trucks) could be constructed. There the gear changes continuously, inversely proportional to the power needed. For neural networks this would mean the learning rate would automatically decrease when average learning rate would stop descending but for this implementation of backpropagation such an algorithm was not (yet) implemented.

I applied this method to 3 configuration, which proved the most promising results during the training of the permutation of parameters; all are with 200 neurons in the hidden layer:

1 greyscale images with high momentum (0.8) and high learning rate (0.1)

2 binary images with a microcalcification with minimal size 10 and removing extra microcalcifications, no momentum and relatively low learning rate (0.01)
3 cut (9x9) binary images with a microcalcification with minimal size 4, no momentum and relatively low learning rate (0.01)

1) Learning he greyscale images started of well, Figure V-5 shows a continuously decreasing error value for the same learning rate even after 1000 iterations. The learning rate is stepwise decreased to 0.0001 (lowest learning rate possible without to much rounding errors). At this learning rate it stops at an error value of 0.446, not changing for 1000 iterations. Changing momentum or eta have no or reverse effect. The value reached is still to high (only 15 percent of input patterns is correctly classified). Closer examination of network weights show high weight values. This means weight values are to high to still be changed leading to network paralysis.

Table V-11: Final result greyscale images

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und	rmsbad	rmsund
200	0.8	0.1	2100	0.446	1.695	0.857	666	15.179	2.670	82.151	1.9	83.1

2) The development of the error value is extremely promising reaching the lowest values after only 77 iterations (See Table V-7). Strange is that the absence of momentum improves the decrease of the error value. Slowly decreasing the eta (among others at 77, 200 and 300 iterations, see Figure V-7) pushes the learning rate to the precision value (default 0.1), the error value where the network stops learning. Lowering the precision value and letting the network work for 10000 iterations finally gives a acceptable value for the error (See Figure V-8). This leaves a trained network ready to be tested.

Table V-12: Final result binary images with removal and size 10.

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und
200	0.0	0.0001	10000	0.081	1.464	0.241	261	87.421	0.629	11.950

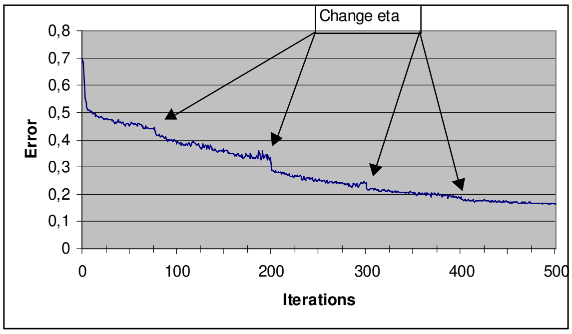

Figure V-7: Error function showing alterations in eta

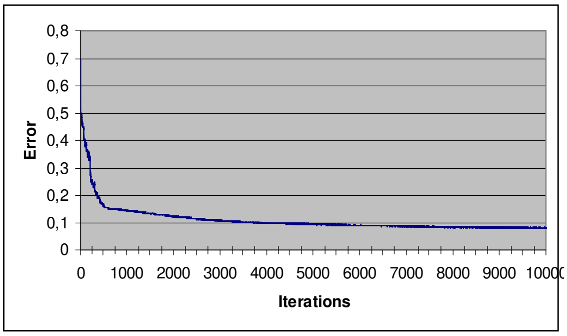

Figure V-8: Error function of whole training process.

3) The last configuration to be learned has reduced size of the image and removal of all images with microcalcifications smaller then 4. This decreases the number of input vectors and the number of input neurons thereby decreasing learning time. The same process happens, which happened for the first configuration. The learning process starts with a gradual lowering of the error value with improvements to this descend by the lowering of the learning rate. But after some time it reaches a certain value where change of learning rate or momentum has no or negative effect. The end error value is still to high for training.

Table V-13: Final result of binary 9x9 images with size 4.

Neu	Mom	Eta	Iter	err	maxerr	averr	pat	%ok	%bad	%und
200	0	0.0001	818	0.264	1.669	0.562	30	51.871	1.548	46.581

7.Testing

After training the moment of truth arrived, feeding the test set to the network and see if it really learned something. Only one configuration reached sufficiently low error rates (<0.1) to be tested. This is the configuration with the 15x15 binary images with minimum microcalcification size 10 and removal of extra calcifications, the classification results of this network on the test set can be seen in the confusion matrix (See Table V-14).

error	maxerror	averror	percok	percbad	percun
0.9264	1.7999	1.1075	30.380	44.620	25.000

Table V-14: Confusion matrix with low delta.

	0	1	2	NOT
0:	7	29	15	11
1:	33	72	31	42
2:	6	27	17	26

To improve these results one can change the **confusion_delta** parameter to its maximum value, thereby forcing the network to choose a class for its output. This modification showed little improvement. Almost all the formerly **undetermined** classified vectors were now classified **bad.** Only 8% of the 25% were classified as good (See Table V-15).

Error	maxerror	averror	percok	percbad	percun
0.9264	1.7999	1.1075	38.608	61.392	0.000

Table V-15: Confusion matrix with high delta.

	0	1	2	NOT
0:	10	33	19	0
1:	45	87	46	0
2:	13	38	25	0

The results are extremely poor. The network does not score higher then a process producing random classifications would do. Of the applied input vectors from the test set ca 30 % as classified good, ca 45 % wrong and the remaining could not be classified. The results proved suspicions, which had surfaced during training. If the network actually is able to learn from this dataset, it is probably because it memorised the whole set instead of learning the characteristics of the I/O-function. The reduced number of input vectors makes it easier to store.

8.Analysis and conclusion

Although the dataset contains the correct information - input vectors with their training pairs - it apparently does not have useful general features to be learned by backpropagation. The dataset is very hard to learn, only one configuration succeeded in reaching the desired error value and that particular configuration tested very poor. Without alterations to the dataset future classification by backpropagation seems highly unlikely. These alterations are the resolution of the images, the independent classification of the individual microcalcifications per breast and the number of cases in the dataset.

The first alteration, improving the resolution of the images, is needed because the extracted images are a very small part off the real images (the mammograms). The microcalcifications are so small that some are only pixel big. At these sizes it is virtually impossible to extract information about the features (like morphology). The resolution is so low that a little bit of noise or an approximation error can push the pixel over the threshold of the segmentation program thereby showing an incorrect form of the microcalcification. To throw away all calcifications smaller then a certain size is a way to subdue image distortion caused by noise but it cuts down on the (already small) number of input vectors. At the Radboud hospital they already have such a dataset but it came too late to influence this project.

Another problem encountered is the false classification of the individual calcifications. If the classification is incorrect to begin with then how can one expect backpropagation to learn the correct classification? This problem is caused by the fact that the diagnosis is determined by the pathologist on the whole breast. This means that for all, say 100, found calcifications the same classification is used. Best would be if only the ones with significant features would be selected for the training set, but looking at this dataset there were numerous identical binary images differently classified. Problem is that it would require knowledge about the exact features for a certain type a microcalcification to program the segmentation program to select the significant calcifications and if such a program would exist what use would the neural network classifier still have? After all the main feature of neural networks is to be able to work in absence of such a theory. The incorrect classification is one of the reasons future work might involve unsupervised networks such as Kohonen's self-organising maps. They find their own classification and with a bit of luck and a lot of trials the map may extract the wanted pattern classification.

The last problem is a problem that plagues most applications of neural networks; the dataset is not big enough. To let a neural network generalise enough examples must be given otherwise it will not learn. The whole set contained 2622 classified microcalcifications from only 56 cases. Compared to the number of input neurons and the possible input values it is only a very small subset. It got even worse after the modifications, which removed too small calcifications from the set. In the case where it did seem to learn something, the dataset was probably so small that the network just stored all input vectors in its memory.

All and all it can be concluded that backpropagation could not learn the desired I/O-function from the given dataset. I would rather have had a positive result but derive comfort from the fact a negative result is a result as well.

Chapter VI: Conclusions

The main purpose for computers is automating human tasks. Originally this was only simple calculation but present applications can handle more and more complex tasks. Artificial neural networks are a relatively new form of automation and are used to automate tasks, which were long thought to be exclusively human. Mimicking the operation of the human neurons they automate specific mental operations with a moderate success. Networks have been created which seem to be able to learn, abstract or generalise. They are still a long way from being real thinking computers but they perform rather well at tasks as pattern recognition

Backpropagation is one type of artificial neural networks. It has layers of neurons that gather, process and classify information. It is a supervised learning method, i.e. it needs a dataset with input vectors and corresponding output vectors. Backpropagation learns by propagating the difference between the target vector and output vector back into the network. It overcomes certain problems of classic perceptron networks but has its own problems like network paralysis and variable parameters. Backpropagation is currently the most used neural network classifier.

The dataset used in this research project originates from the radiological research of breast cancer. This research has a high social relevance because breast cancer is nowadays one of the most common and dangerous diseases among women. Breast cancer can be cured easier if it is found in early stage of development. Mammographic screening is a way to detect breast cancer in a early stage. X-ray pictures are taken from the breast and examined for signs of malignant tumours. Microcalcifications are small calcified pieces of tissue which can be found around malignant tumours but can also have other origins. In this research project it has been tried to distinguish the calcifications signifying malignant tumours from other calcifications.

The mammograms in the dataset were first provided with a classification by a radiologist. Next a segmentation program extracted the microcalcifications. The calcifications together with their classification were then converted to a data for the backpropagation network. Half of them were used for classification, half of them were used for testing. To try to learn the dataset several parameters were altered to improve learning results. Also a number of modifications to the dataset were tried to improve results. All these modifications and alterations finally showed one thing. Backpropagation can not classify the current dataset.

Analysis of the results showed several possible reasons why the dataset could not be learned. The dataset was not correctly classified making it hard for a supervised algorithm as backpropagation to learn. Furthermore the dataset was to small, containing not enough information for the network to distinguish general features. The last reason was the resolution of the pictures, good enough to look at features as whole tumours or clusters of microcalcifications but not good enough to detect features in microcalcifications.

1.Future work

There are several areas where the research into classification of microcalcifications may advance. For one the dataset may be improved. The images can be taken with a higher resolution causing the microcalcifications to have more graphic detail and as such more extractable information. The classification of the test set can be improved by classifying the individual classifications not as on the whole breast but on single calcifications. More cases would also be very convenient to exclude coincidental correlations in the dataset. All these advises are meant for the radiologists not the neural network implementer and are probably acknowledged ideas. But I as an implementer just state my observation that dataset is lacking in these aspects.

Backpropagation itself could also be improved. An automatic adjustment of the learning rate would improve the autonomy of the learning process. This way backpropagation would not constantly require monitoring by the controller. Another thing that bothers me is the ignorance of backpropagation that it is dealing with 2-dimensional data. The network just receives a one-dimensional input vector and has as such for instance no information, which pixels are beneath each other. This could be improved by not connecting all the neurons in the input layer with those in the succeeding hidden layer. A neuron in the hidden layer would only receive information from input neurons, which lie in the each other's neighbourhood. This implementation would require a severe modification of the backpropagation algorithm because the calculation of the output value and the backpropagation of error would change.

As a last addition I would recommend unsupervised learning. This originally would be part of this thesis, but was left out due to time constraints. Unsupervised learning can overcome the problem of a faulty classification. The network itself may be able to recognise certain features and make its own classification. One could then test all the classifications in one breast and see in which classes they score the best. This global classification could then be compared to the current classification and checked for similarities.

A last point is leaving the subject of classifying individual classifications altogether. The research at the Radboud is already concentrating on the clustering of classifications. Maybe neural networks can be of help here because for classification of clusters it is sure there is information to be found, after all this is the way radiologist originally make their diagnosis

Thesaurus

carcinoma a malignant tumour of epithelial origin.

comedo (glutton =) one that has a great capacity for accepting or enduring something.

cyst a closed sac having a distinct membrane and developing abnormally in a cavity or structure of the body.

epithelium a membranous cellular tissue that covers a free surface or lines a tube or cavity of an animal body and serves especially to enclose and protect the other parts of the body, to produce secretions and excretions, and to function in assimilation.

exocrine gland external organ which secretes fluids.

fibroadenoma a benign, fibrous tumour of a glandular structure or of glandular origin.

fibrous fibre-like.

histology tissue structure or organisation.

in situ in the natural or original position or place.

ligament a tough band of tissue connecting the articular extremities of bones or supporting an organ in place.

metastasis transfer of a disease-producing agency from the site of disease to another part of the body.

necrosis usually localised death of living tissue.

tumor/tumour an abnormal benign or malignant mass of tissue that is not inflammatory, arises without obvious cause from cells of pre-existent tissue, and possesses no physiologic function.

Appendix A: Checklist

History

Self-history : Contralateral breast cancer?(high risk!)
How many pregnancies ? (nulliparity means a higher risk !)
Previous biopsies:
If so, check previous histologic findings! An unequivocal microcalcification cluster must be interpreted differently if the patient had previously had a true proliferative cystic disease possibly with atypia, than if the lesion had been completely innocent.
Other trauma ?
Iatrogenic trauma ?(e.g., plastic surgery?)

Clinical Examination

Palpation:
If a palpable mass is found, is it assignable to the visible microcalcification cluster? For example, a patient referred for evaluation of a simple cyst may be found also to have a microcalcification cluster suggesting intraductal carcinoma.
Nipple changes ?
Any kind of eczematous changes, even minimal exfoliation, may raise suspicion of Paget's disease, in which case even minimal or uncharacteristic microcalcifications will assume importance.
Discharge:
Nipple discharge is of little help diagnostically, because it occurs in both benign and malignant conditions. Even a bloody discharge is not pathognomonic for intraductal carcinoma. Galactography can be used to determine whether the questionable microcalcification cluster is related to the secreting duct.

Mammographic Interpretation

Mammograms are routinely taken and evaluated on two planes. (Hereafter, lack of reference to the second plane means that that view contained no useful information.)
Answer the following questions:

1. Is only one type of microcalcification seen, or are several different patterns present together?
(e.g. , fibroadenoma and sebaceous gland, fibroadenoma and carcinoma, etc.). The examiner is cautioned not to overlook an intraductal carcinoma because of an easily diagnosed coexisting lesion (e. g., a "classic" calcified fibroadenoma)!

2. Where are the microcalcifications located?
Extramammary localisation? If so, is there an artefact on the film? on the skin? intracutaneous?
Intramammary localisation ?

Is an intraductal localisation suggested by the cluster shape (see p. 91)?
What is the appearance of the cluster contours (see p. 95)?
If the calcifications are not intraductal, are they:
a) interlobular?(cluster shape, see p. 42, 59)
b) inside a fibroadenoma ?(cluster shape, see p.146)
c) in connective tissue? fatty tissue?(history of trauma or previous surgery?) (see p.157)
d) intra-arterial?(see p. 176)

Mixed localisation? For example, intraductal and interlobular (as with coexisting milk of calcium cysts and carcinoma), intraductal and interductal (interstitial) connective tissue localisation (e.g., calcified secretion and liponecrotic cysts; plasma cell mastitis), or intraductal and intracutaneous localisation (e.g., carcinoma and
calcified sebaceous glands).

3. What is the shape of the microcalcifications ?

Monomorphous? Polymorphous?
If monomorphous:
a) Are they in both views? ("teacup sign"? see p.51)
b) Is septation present? (see p. 42)
If polymorphous:
a) Does the polymorphism increase in accordance with the number of microcalcifications? Caution: The more extensive a ductal carcinoma, the larger its comedo component and thus the greater the polymorphism (see p.108). Even minimal polymorphism is suspicious!
b) Is the cluster shape consistent with an intraductal process? Consider that polymorphous microcalcifications are found not only in intraductal lesions and in carcinoma but also in fibroadenoma (see p.152), in sclerosing adenosis (see p.48), in obliterative comedomastitis (see p.133) and, rarely, in fat necrosis (see p.173).
Remember also that the cluster shape is sometimes difficult to evaluate in very small lesions (3-4mm in size).

4. Is the change unilateral or bilateral?

If homogeneous, predominantly punctuate, slightly polymorphous microcalcifications are present bilaterally, an intraductal carcinoma is unlikely. (The author has not personally seen a case of simultaneously occurring bilateral duct carcinomas with associated microcalcifications, though he has seen extensive bilateral papillomatosis).

5. How should microcalcifications adjacent to a soft-tissue shadow be interpreted?

When microcalcifications are seen in the vicinity of a soft-tissue shadow suggestive of carcinoma, there is a tendency to draw a connection between the two radiographic signs. But soft-tissue shadows typical of carcinoma can both facilitate and hamper the differential diagnosis of microcalcifications. A microcalcification cluster in proximity to a scirrhus, for example, is not necessarily malignant in nature. Neither does a confirmed carcinoma on one side necessarily mean that a microcalcification cluster in the opposite breast also signifies carcinoma. Conversely, an incipient carcinoma marked by a small microcalcification cluster may be discovered incidentally in the vicinity of a wholly benign lesion.

A rounded, slightly lobulated shadow with smooth margins and a "halo sign" (fatty margin) establishes the diagnosis of a calcifying fibroadenoma even if the microcalcifications visible within this shadow are polymorphous. Microcalcifications in cyst walls are infrequent, and they are usually localised at the periphery of the round, smooth-bordered shadow on one plane. They are easily identified as cyst-wall calcifications on the pneumocystogram.

Appendix B: De PowerXplorer

Figure B-1 shows the 4 four partitions (cabinets) of the PowerXplorer. Minimum use is restricted to 1 cabinet per user and a maximum of 4 (16 processors).

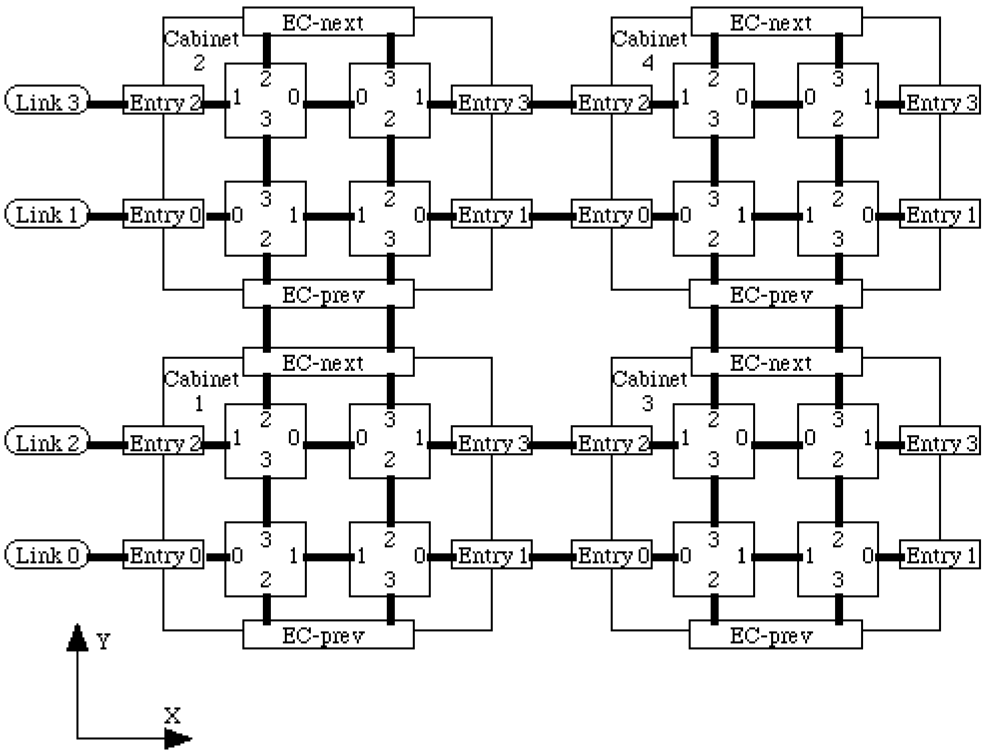

Figure B-1: Topography of PowerXplorer

The links are via a BBK-SCSI box connected to POSEIDON, a SUN SPARCstation 20 workstation with two 75 Mhz SuperSPARC-II CPU's and 64 Mbyte memory

Bibliography

Bon Martens, M.H.J van, Verbeek, A.L.M, Peeters, P.H.M, Luning, P, Werre, J. M. 1990, Een overzicht van de epidemologie naar borstkanker; sterftereductie in binnen en buitenland. In *Nederlands tijdschrift voor Geneeskunde,* 1990; 134, nr46: 2240-2245.

Chelberg, D. M., Hsu J., Babbs C. F., Pizlo, Z., Delp E. J..1994 Digital Stereomammography, *Proceedings of the 2nd International Workshop on Digital Mammography*, Excepta Medica International Congress Series 1069, July 1994, pp 181-190,Elsevier Science, York, England.]

Citoler P. 1978. Microcalcifications of the breast. In: Grundmann E, Beck L (eds) *Early diagnosis of breast cancer*. Fisher, Stutgart, p 113.

Dennett, D. C. 1991. *Consciousness explained*, Boston, Mass:Little Brown.

Frieling, J.T.M 1992. *Differentiation of microcalcifications in digital and conventional mammography. Department of diagnostic radiology university hospital*. Nijmegen.

Grossberg, S. 1973. Contour enhancement, short time memory, and consistencies in reverberating neural networks. *Studies in Applied Mathematics* 52:217, 257.

Doyle R. 1995. *Deaths Caused by Breast Cancer, by County*. Scientific American. October

Hebb, D. O. 1961. *Organization of Behavior*, New York, Science Editions.

Karsemeijer N. 1993 Adaptive Noise Equalization and Recognition of Microcalcification Clusters in Mammograms, In *International Journal of Pattern Recognition and Artificial Intelligence*, Vol. 7, No 6 (1993) 1357-1376, World Scientific Publishing Company

Keys, Robert G 1981. Cubic Convolution Interpolation for Digital Image Processing IEEE *Transactions on acoustics, speech and signal processing*, Vol. ASSP-29. No. 6, December

Kohonen, T. 1988. *Self-organization and associative memory*. 2d ed. New York: Spinger-Verlag.

Lanyi, M 1986. *Diagnostiek und Differentialdiagnostiek der Mammaverkalkungen.* Berlin. Springer-Verlag.

Le Gal M, Durand JC, Laurent M, Pellier D 1976. *Condute à tenir devant une mammographie révelatrice de microcalcifications groupeés, sans lumeur palpable*. Nouv Presse Med 5:1

Le Gal M, Chavanne G, Pellier D. 1984. *Valeur diagnostique des microcalcifications groupeés déconvertes par mammagraphies*. Bull. Cancer. Paris 71/1:57

Minsky, M. L., and Papert, S. 1969. *Perceptrons*. Cambridge, Ma: MIT Press.

Parker, D. B, 1982. *Learning Logic*. Invention Report S81-64, File 1, Office of Technology Licensing , Stanford University, Stanford, CA.

Rumelhart, D. E., Hinton, G. e., and Williams, R. J. 1986 Learning internal representations by error propagation. In *Parallel distributed processing*, vol. 1, pp. 318-62. Cambridge, MA:MIT Press.

Salomon A .1913. *Beitrage zur Pathologie und Kliniek der Mammakarzinome*. ArCH Klin Chir 103: 573

Sejnowski, T. J., and Rosenberg, C. R. 1987. An improved three-layer backpropagation algorithm. In *Proceedings of the IEEE First International Conference on Neural Networks*, eds. M Caudill and C. Butler. San Diego, CA:SOS Printing.

Stegner HE, Pape C. 1972. *Beitrag zur Feinstructur der sog. Mikrokalzifikation in Mammatumoren.* Lbl Allg Path Anat 115:106.

Wassermann, P. D. 1988a Combined backpropagation/Cauchy machine. *Proceedings of the International Neural Network Society*. New York: Pergamon Press.

Wassermann, P. D. 1989. *Neural Networks*. New York. Van Nostran Reinhold

Werbos, P. J. 1974 *Beyond regression: New tools for prediction and analysis in the behavioral sciences.* Masters Thesis, Harvard University.