LU TP 03-05

Comparing prognostic markers for metastases in breast cancer using artificial neural networks

Master thesis by Cecilia Ritz

Thesis advisor: Patrik Edén

Department of Theoretical Physics, Lund University,
Sölvegatan 14A,
S-223 62 Lund, Sweden

Abstract

An important task of breast cancer treatment is to accurately diagnose the risk of metastasis. In this thesis artificial neural networks are used to compare the predictive power of conventional clinical observables and gene expression profiles of breast cancer tumours. Publicly available data from a previous study are used. In contrast to that study, we do not find that gene expression data outperforms conventional clinical observables.

Contents

1 Introduction

A metastasis is a secondary tumour caused by tumour cells migrating from the primary tumour. Metastasis can occur several years after the primary tumour has been removed, and is often lethal. Treatments against metastases exist, but are usually very painful and not without risk. An important task of cancer treatment today is therefore to accurately diagnose the risk of metstasis, to optimize the metastasis-inhibiting treatment.

Many different observables, such as the size of the primary tumour and the age of the patient, are used to achieve this goal. In this way, the clinical outcome can be estimated but not completely predicted. Some low-risk patients unexpectedly develop metastases, and some high-risk patients that for other reasons are untreated, unexpectedly remain metastasis free for a long time.

With the development of microarray techniques (described more below), the gene expression profile of the primary tumour could be added to the list of prognostic tools for clinical outcome. According to van't Veer *et al.*[1] and van de Vijver *et al.*[2], this significantly improves the chances for a correct prognosis.

The experimental data used in [1] is publicly available. This data consists of gene expression profiles and clinical observables for 97 samples of primary breast cancer tumours. 51 patients remained free from metastases for at least 5 years, metastasis negative, and 46 of the cancer patients developed metastases within 5 years, metastasis positive. All patients were under 55 years of age and were lymph node negative *i.e.,* they had no tumour cells in local lymph nodes.

In this study, we analyse the data using artificial neural networks (ANNs), to further compare gene expression profile and the conventional clinical observables as prognostic markers. Due to the great amount of data, 25000 genes measured for each tumour, and the possibly nonlinear correlations between them, it is suitable to use artificial neural networks for classification of the tumours. The flexibility of ANNs makes it possible to use both clinical observables and gene expression profiles in prediction of metastases, in combination or separately. This allows for comparison of the predictive power between the two types of data *i.e.,* which one contains most information about metastases.

2 DNA microarray analysis

DNA microarray analysis is an experimental method to simultaneously measure the gene activity for thousands of genes.

The genes are responsible for regulating the protein composition of the cell. Each gene codes a protein. The protein production units of the cell, the ribosomes, have

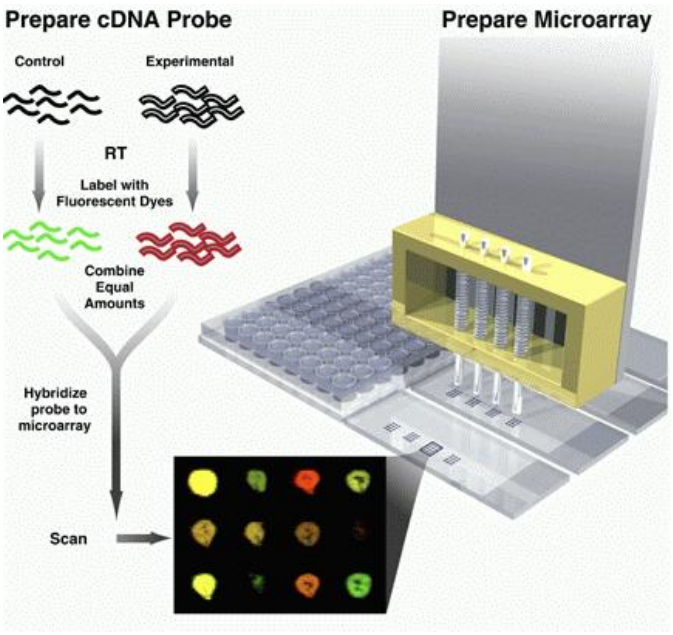

Figure 1: *Microarray technology.*

no access to the information in the DNA. For the protein to be created the gene has to be transcribed into a temporary copy, messenger RNA (mRNA), that carry the information to the ribosomes. While being transcribed, a gene is said to be 'active' or 'expressed'. The amount of corresponding mRNA is thus an indication of the gene expression.

A microarray is an array of nucleotide sequences. One spot on the array contains a DNA sequence that is unique for one gene. The RNA is extracted from the cells and during reverse transcription* labeled with a flourescent molecule (red). It is mixed with a reference sample (green labeled) and flushed over the microarray. All cDNA molecules from one gene (with both labels) will hybridize to one spot on the array. The array is then scanned to measure the flourescent intensities in each spot. Figure 1 schematiclly illustrates the principles of this experimental technique.

The total intensity of a spot differs from gene to gene and from array to array. This variation depends not only on the gene activity but also on unavoidable variations in *e.g.,* spot size and occurence of dust on the array. The reference samples is used to obtain a ratio of intensities where many such systematic errors are eliminated.

The ratio of red and green intensities gives information on wich genes are more active (up-regulated) or less active (down-regulated) in the red sample compared to the reference sample. Using the same reference on all microarrays means that the gene expressions of different samples may be indirectly compared. The logarithm of the ratio is used to make the up- and down-regulations with respect to the reference

*Reverse transcription is a process where a cDNA copy is produced from a RNA template.

sample equally important. If the mRNA levels are the same in both samples the log(ratio) is equal to zero.

To correct for different incorporation rates of the dyes or unequal amount of RNA in the two samples, the distribution of log(ratio) is normalized [4]. The public data used here is already normalized and the normalization is not further investigated here.

3 Data

The data used in this study consists of approximately 25000 gene expressions and 7 clinical observables measured for each tumour.

The clinical observables used are the age of the patient, the size of the tumour, angioinvasion, ERp, PRp, tumour grade and lymphocytic infiltrate.

Angioinvasion is whether tumour cells have invaded blood or lymph vessels or not, and lymphocytic infiltrate is the presence of lymfocytes in the tumour, both binary variables. The levels of receptors of the female hormones oestrogen and progesteron in the tumour cells were measured with immunohistochemical staining and the percentage of nuclei that show staining is reported, ERp and PRp. The tumour is ER/PR negative if ERp/PRp $\leq 10\%$ and positive otherwise. The tumour grade is a parameter of cancer cell abnormality, high, low or medium. It includes tissue organization of the tumour, nuclear shape and size of the tumour cells and the growth rate of the tumour.

In [1], two hybridizations were carried out using the RNA extracted from one tumour. One microarray with red tumour sample and green reference sample and one with the reversed dye assignments. A pool of RNA from all tumours were used as a reference sample. This implies that genes with a mRNA abundance equal to the average mRNA abundance over all tumour samples will get a log(ratio) equal to zero.

3.1 Error model

The errors in the measurements of gene expression is estimated using an error model described in [5].

The error model relies on the duplicate measurements made for each tumour. The two observations of log(ratio) is combined in a weighted average $\bar{x}$. The weight w_i is calculated from the uncertainties in the observations σ_i.

$$\bar{x} = \frac{w_1 x_1 + w_2 x_2}{w_1 + w_2}, \quad \text{where} \quad w_i = \frac{1}{\sigma_i^2} \tag{1}$$

σ_i is defined as the ratio between the measurement of log(ratio) and its significance. The significance of an observation increases with spot intensity and decreases with background intensity. The errors σ_i are then combined to a total error of $\bar{x}$ that depends on the agreement between the two observations. The total significance, X, is ultimately used in the classifiers.

$$X = \frac{\bar{x}}{\sigma_{\bar{x}}} \tag{2}$$

Since X is found to have approximately a standard normal distribution the probability P for an observation of magnitude $|X|$ could be calculated using equation

$$P = \frac{2}{\sqrt{2\pi}} \int_{|X|}^{\infty} e^{-t^2/2} \mathrm{d}t. \tag{3}$$

Present in the data files are the weighted average of log(ratio) and the P-values but not X. Therefore it is reconstructed using the matlab function erfcinv (inverse of complementary error function). The sign of X is determined by the sign of log(ratio).

3.2 Gene filtering

The filtering step removes the genes without significant regulation, *i.e.,* genes that do not have a large enough intensity ratio in sufficiently many samples. This is done because with a huge number of genes one would always find many weakly expressed genes that correlate with the metastases class purely by chance and influence the classification without having any predictive power.

The first condition the gene has to fulfill in order to pass the filtering is to have a P-value less than 0.01 in more than 3 experiments.

The second condition is that the gene should have a (magnitude of) log(ratio) that exceeds log(2) in more than three experiments.

In the gene filtering performed here the two conditions were not required to be fulfilled for the same three experiments. A third demand not required in [1] but necessary when the gene expressions are used for training the ANNs, is that the gene is measured in all experiments.

3.3 Gene selection

The performance of a classification method depends to a great extent on the ability to find the signal in the input data. With gene expression data the problem is to select the genes that indicates metastasis.

The genes with a significant difference in gene expression between the two classes are assumed to be the genes important for the classification. Thus the ranking of genes are based on the Pearson correlation $C_P$between the gene expressions and the metastasis class.

$$C_P = \frac{\sum_i (m_i - \bar{m})(g_i - \bar{g})}{\sqrt{\sum_i (m_i - \bar{m})^2 \sum_i (g_i - \bar{g})^2}} \tag{4}$$

g_i is the gene expression of tumour i and $m_i = 1$ if the tumour developed a metastasis and $m_i = 0$ otherwise.

The genes with the highest magnitude of Pearson correlation coefficients are then used in the classifier.

In [1] a permutation test was performed to find the value of $|C_P|$ for which the gene could be considered to be significantly associated with metastases. The class labels were randomly permuted and for each permutation the Pearson correlation for all genes was computed. By comparing the Pearson correlation distribution between the metastasis class and a random class the probability for a gene to have a Pearson correlation greater than 0.3 by chance was estimated to be sufficiently small.

4 Artificial neural networks

An artificial neural network (ANN) is a directed graph where each edge is assigned a weight. A subset of the nodes are input nodes and another subset (or the same) are output nodes. The rest are called the hidden nodes. The value of a node is the weighted sum over the output from all nodes with an incoming edge. The output from one node is a function of this sum.

The networks used here are feed forward networks. The only connections are from the input nodes to the hidden nodes and from the hidden nodes to the output nodes, see figure 2.

Output from the artificial neural network is a function of the input $\bar{x} = (x_1, ..., x_N)$ (and a threshold $x_0 = 1$). In the case of a classification problem with two classes one output is sufficient. The output o from a network with no hidden nodes and one output node (the perceptron) is given by

$$o = \varphi(\sum_{i=0}^{N} w_i x_i). \tag{5}$$

For a network with M nodes in the hidden layer:

$$o = \varphi(\textstyle\sum_{j=0}^{M} w_j h_j) \tag{6}$$

$$h_j = \tilde{\varphi}(\textstyle\sum_{i=0}^{N} \tilde{w}_{ij} x_i) \tag{7}$$

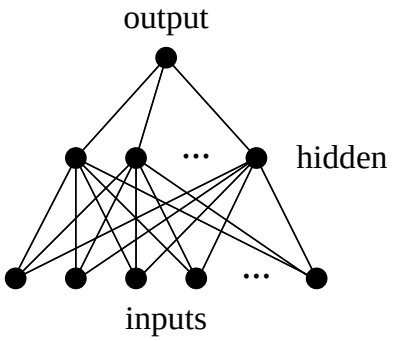

Figure 2: *multilayer perceptron*

The weights are determined by minimizing the error-function E between the network output o and target t, a sum over K measurements.

$$E = \sum_{i=1}^{K} (o_i - t_i)^2 \tag{8}$$

This is called training the network.

4.1 Decision surface

The output activation function φ is sigmoidal and varies for example between -1 and 1. Then $o = 0$ defines a decision surface in input space that hopefully can separate the samples of the two classes. In the case of the perceptron:

$$o = 0 \quad \Leftrightarrow \quad \sum_{i=0}^{N} w_i x_i = 0,$$

where the rightmost equation defines a hyperplane. Therefore the perceptron is a linear classifier and can only solve linearly separable problems. Often this approximation is good enough because several hidden nodes will make the network too flexible and cause overtraining (see below). But all problems are not linearly separable and in these cases an ANN with hidden nodes might be a better choice. Every node in the hidden layer define a decision plane and the hidden nodes combine them to a non-linear decision surface.

4.2 Overtraining

The flexibility of the ANNs that makes them so useful in many applications also causes problems during training. While the net learns the features of the classification problem there is a risk that it will learn how to classify the trainingsamples

at the cost of decreased performance on a blindtest, especially if the trainingsamples are not completely representative for the problem or if there is noise in the measurements.

The number of parameters in the ANN to be determined by the training has to be less than the number of trainingsamples but sufficiently many to solve the problem. For instance hidden neurons can create a non-linear decision surface to solve a not linerly separable problem, but a net with more parameters in form of weights will need more trainingsamples.

One way to detect overtraining is to validate the performance of the network. This is done using a K-fold crossvalidation scheme. The trainingsamples are divided into K groups. One of the groups is left out in turn to serve as a validation test of the net that is trained on the remaining $K - 1$ groups. The validation result can be used to choose a suitable number of training epochs so that the training stops before the validation error starts to increase due to overtraining. It can also be used for selection of the best network architecture.

To some extent, overtraining can be avoided by pruning of the network. Pruning means that unnecessary edges are removed from the network during training. This can be done by adding a term to the error function that will encourage the weights that have little influence on the training result to approach zero.

4.3 Committee

Creating a committee of networks trained on different subsets of input space is a way to avoid overtraining and to make the training faster. The members of the committee are trained on different subsets of the training samples. Accordingly they become specialized on different features of the problem. The blindtest result depends on how well the individual nets can classify the blindtest but also on the spread of the committee output. If the committee is too homogenous the blindtest result will most likely improve if only one net is trained on all trainingsamples. The ideal committee has a good blindtest performance of the individual nets and at the same time a large variance of the output.

4.4 Principal component analysis

The purpose of principal component analysis (PCA) is to reduce the number of input variables. When using gene expression as input to the ANN the number of genes is much larger than the number of traingsamples available.

Each sample is described by a vector of gene expressions, coordinates in gene space. The PCA finds the linear combination of genes that gives the direction in gene space

where the variance of the samples is largest. This is the first principal component. The second principal component is the linear combination that contains the largest variance under the condition that it is orthogonal to the first. This continues until the number of components is equal the number of samples.

The result of the PCA is a rotation of the original coordinate system such that all information is in the principal components.

The first principal component has the largest variance over the samples but it is not necessarily the most important variable for the classification. When the variation in the components is less than the uncertainties in the measurements they cannot contribute to the training of the network, and are not used as inputs.

5 Classification methods

Creating a good classifier requires many data points with known class assignments. In the case of ANNs all the parameters of the network is determined by minimizing the difference between the network output and the known class of the input, the target.

To know if the classifier can correctly classify a new sample, it is applied to a blindtest that has not been involved in training or validation of the network. The performance of the ANN increases with more trainingsamples but the number of trainingsamples is limited, so to use as many trainingsamples as possible and at the same time perform many blindtests of the classifier a leave-one-out procedure is followed. Each sample is in turn left out to serve as a blindtest of the classifier derived from the remaining samples.

The gene selection has to be redone for every left out sample. If the blindtest was allowed to influence the gene selection, the genes that correctly classify the blindtest would be favoured. An information leak has occurred and the blindtest would no longer be a blindtest. The result in section 7 shows that this information leak is important.

5.1 ANN settings

The networks are trained with a program called `backp` [3] because of the backpropagation method used in minimization of the error function, equation (8).

Activation function in the hidden layers is $\tilde{\varphi}(x) = tanh(x)$ and the output activation function is

$$\varphi(x) = \frac{1}{1 + e^{-2x}}. \tag{9}$$

Input to the network is either clinical observables or principal components of gene expression profiles or both. In the case of gene expression profiles one could either choose to use the first principal components or to rank order them and choose the principal components with the highest rank as input. To rank the principal components a network is trained using all principal components, in wich the sample variance is large enough, as inputs. Every input is then assigned a rank based on the impact on the network output when the input is deleted.

The best network design is determined using a 3-fold cross-validation procedure. For each design a committee of nets is created. All samples except for the blindtest are divided into three groups. 3 nets are trained using two of these groups as trainingsamples and the third group of samples as a validation set. The partitioning of the samples in three groups with equal size is random and repeated 20 times. This results in a committee with 60 members with validation results recorded. The committee with the best validation result is then applied to the blindtest.

The pool of designs include ANNs with 2 or 4 principal components as input and if these should be the first principal components or chosen among the 20 first principal components where the variation is expected to be larger than the background noise. There is also a choice between linear ANN and an ANN with 2 or 3 hidden nodes and pruning or no pruning of the network.

This results in 24 different designs for the networks trained on gene expressions and 6 in networks trained using clinical observables where the number of input nodes is fixed.

In a preliminary study, different ANN architectures and learning parameters were tested, and the learning curves investigated. Based on the results, ranges for different parameters of the ANN were determined. For some parameters, like number of epochs and learning rate, a single value was found to perform well under all circumstances, and these parameters were held fixed in the following.

Reducing the range of parameters, and fixing some of them, results in an necessary reduction of computation time. In principle, it also introduces an information leak, but, since many essential degrees of freedom remain to optimize in the following analysis, the biasing effect is not worrying.

5.2 Cone algorithm

This classification method is based on the correlation between the average gene expression profile over the metastasis negative tumours and the gene expression profile of the tumour that is classified. The measure of correlation is

$$C = \frac{\sum_i g_i t_i}{\sqrt{\sum_i g_i^2 \sum_i t_i^2}}. \tag{10}$$

Here g_i is the expression value of gene i in the left out sample and t_i is the average of gene expression i over metastasis negative samples. If the correlation in equation (10) is high enough the tumour is classified as metastasis negative.

The average gene expression is not subtracted before calculating this correlation coefficient, in contrast to the Pearson correlation coefficient, because the sign of a gene expression is important. If the average gene expression is subtracted from the gene expression profile, the gene expression profiles of a tumour with all gene expressions upregulated and a tumour with all gene expressions downregulated could be completly correlated.

C is the scalarproduct between the average gene expression profile and the blindtest divided by their length in gene space. This is by definition the cosine of an angle in a multidimensional space. Therefore the choice of name of the method, because the requirement that the correlation is larger than a number between 0 and 1 defines a hypercone in gene space. If the blindtest lies within this cone it is classified as metastasis negative.

6 Quality of prediction

The performance of a classifier can be measured by how many samples are incorrectly classified in a blindtest. This requires a fixed cutoff that can tell whether samples are assigned to the metastas positive or the metastas negative class.

The output from an ANN could be interpreted as the probability to belong to one of the two classes. To count the errors, it is therefore reasonable to use the cutoff 0.5. In the cone algorithm the cutoff is set to a level where less than 10% of the metastasis positive samples are misclassified.

A misclassification where a metastasis positive tumour is assigned to the metastasis negative class is a false negative result. A false positive result is obtained when a metastasis negative tumour is assigned to the metastasis positive class. In the prediction of metastasis in breast cancer a classifier that produces less false negative than false positive results is preferred.

6.1 Odds ratio

In [1] the success of a prediction is measured using its odds ratio. The odds ratio is the ratio between the odds in favour of a correct classified metastasis positive sample to the odds in favour of a misclassification of a metastasis negative sample, or equivalent, the ratio between the odds in favour of a correct classified metastasis negative sample to the odds in favour of a misclassification of a metastasis positive sample. Either way, a successful prediction gives a large odds ratio, but it depends

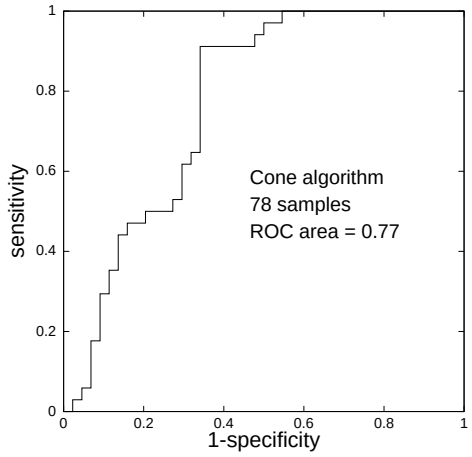

Figure 3: *example ROC curve*

on the chosen cutoff.

6.2 Area under ROC curve

The ROC (Reciever Operating Characteristics) curve shows the result of the prediction at every possible cutoff. 1−specificity is plotted versus sensitivity. Sensitivity is the probability that a positive sample is correctly classified and specificity is the probability that a negative sample is correcly classified. The area under this curve represents a value of the quality of the prediction that is independent of the chosen cutoff. The ROC area could therefore be used for comparison of different methods for prediction.

For a classifier based on random guessing the ROC area would be close to 0.5 and for a perfect classfication the ROC area is equal to 1.

An example of a ROC curve is shown in figure 3. It shows the blindtest result of the cone algorithm with 78 tumours. The ROC area is 0.77, indicating that there is a 77% chance that a randomly selected metastasis negative sample has a larger output than a randomly selected metastasis positive sample.

Here the ROC area is used for comparison of the performance of different classifiers, both different classification methods, cone algorithm versus ANN, and different input data to the ANN, clinical observables versus gene expressions.

A permutation test is used to determine the significance of a difference in ROC area, ΔROC, between two classification results. To compute the ROC area it is sufficient to know the numerical order of the sample output and class labels of the samples. The rank of each sample is swapped between the two results with a probability of 50%. A ROC area difference is computed for the two new permuted results. This

classification methods	78 samples	97 samples	ER positive samples
cone algorithm	0.77	0.76	0.79
ANN gene expression	0.71	0.70	0.68
ANN clinical observables	0.76	0.77	0.83
ANN combination	0.77	0.71	0.83

Table 1: *ROC area for classifications with a new gene selection for each blindtest.*

classification methods	78 samples
cone algorithm	0.85
ANN gene expression	0.81
ANN clinical observables	-
ANN combination	0.87

Table 2: *ROC area for classifications with one gene selection based on 78 tumours.*

is done several times to get a good estimation of the probability that a random permutation of the two results will have a ROC area difference that exceeds ΔROC.

7 Result

The methods of gene filtering and selection described in sections 3.2 and 3.3 are used in [1] where the cone algorithm for prediction of metastasis in breast cancer is introduced. Their results were successfully reproduced and the same method of gene selection is used in the ANN classifier to be able to compare the performance of the two classifers.

After gene filtering, the number of genes is reduced from 25000 to 5200. The filtering was based on 78 of the 97 tumour samples available, in order to be as similar as possible to the analysis in [1]. Some differences appeared, due to our exclusion of genes with missing values. The missing values constraint reduced the number of accepted genes from 5400 to 5200. Among the 231 important genes, found in [1] to have a significant correlation ($|C_P| > 0.3$) to metastasis among the 78 samples, only 7 were excluded due to missing values.

Another possible source of differences between the original data used in [1] and our reconstruction, is small numerical errors in the calculation of the inverse of the complementary funcion, equation 3. The differences are, however, very small. Of the 231 available genes that had $|C_P| > 0.3$ in [1], only 5 had a lower value in our calculation. Only 22 other genes in our gene list got a $|C_P| > 0.3$. All these genes for which the selection differed had a $|C_P|$ very close to 0.3.

Table 1 shows the ROC area calculated from the classification result of different

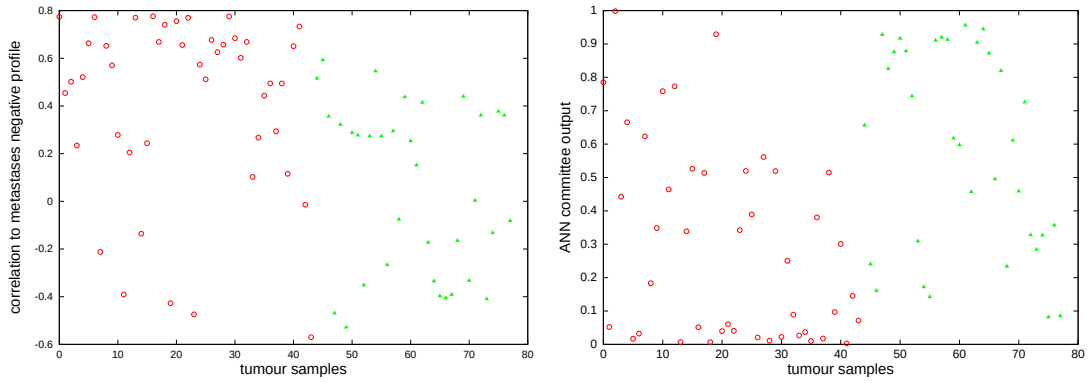

Figure 4: *The blindtest result of the cone algorithm (to the left) and ANN clinical observables(to the right). Circles are metastasis negative samples and triangles are metastasis positive.*

classifiers and different sample selections. The 78 samples are the same as used in the original cone algorithm where the remaining 19 were used as a blindtest of the classifier derived from one gene selection based on all 78 samples. One classification was performed using only the ER positive samples. There were 68 ER positive samples, 41 metastasis negative and 27 metastasis positive.

One network design is chosen for each blindtest. The most popular ANN design with gene expression as input was the net with the two first principal components as input and two hidden nodes, except for the ER positive samples where a linear classifier was more successful. For all ANN classifiers with clinical observables as input (6 or more input nodes) a design with no hidden nodes was preferred.

In the ANN classifier using both gene expression and clinical observables the two first principal components were used as input in addition to the 7 clinical observables. In the classification of the ER positive samples there are only 6 clinical observables.

The P-value of ΔROC between the ANN classification of ER positive samples using gene expressions and the ANN classification of ER positive samples using both gene expression and clinical observables is the only P-value below 0.01 in table 1. The P-values of ΔROC between the cone algorithm and ANN classification using gene expressions are below 0.1 in all sample selections. ΔROC between ANN using gene expressions and ANN using clinical observables from the ER positive samples and the ΔROC between ANN using clinical observables and ANN using both gene expression and clinical observables in combination for the 97 samples have P-values below 0.1.

Table 2 shows the increase in the performance of the classifiers due to the information leak introduced in the gene selection.

8 Discussion

8.1 Main result

We do not confirm the result in [1], that the gene expression profile is superior to classical clinical observables as a prognostic marker for clinical outcome. The performance of the ANN classifier based on clinical observables is very similar to that of the best classifier based on genes alone.

One should note that the different conclusions rely on somewhat different forms of comparisons. We compare ROC areas, and use a permutation-based test to get the significance of ROC area differences. In [1], a gene-expression based classifier (the cone algorithm) and conventional clinical observables are compared using odds ratios between the prediction and the outcome. In that comparison each clinical observable was converted to a binary format before analysis, and obviously some of the available information was lost.

8.2 ANN performance

The cone algorithm outperforms the ANNs when only genes are used. There are two possible explanations.

First, it may be that the poorer performance of ANNs compared to the cone algorithm is related to gene selection. Since the validation samples contributed to gene selection, the risk of overtraining might be underestimated by the validation. To test if this is a relevant problem, a new algorithm, where genes are reselected for each training set, must be tested. With the current use of 60 networks in a committee, that implies a factor of almost 60 in computation time, since the ranking of 5200 genes is one of the most time consuming steps in the calculation.

The results based on 78 samples in table 1 and 2 (new gene selection for each blindtest vs. one single gene selection), show that the expected information leak in gene selection is relevant. The P-values of the difference in ROC area between classifications of the 78 samples in table 1 and 2 are for the cone algorithm and ANN gene expression < 0.0001 and for the ANN combination < 0.01.

When introducing the information leak through the gene selection based on all 78 tumour samples, the two most used network desings chosen by the ANN classifier are the same as for the ANN classifier with a new gene selection for each left out sample. It still falls behind the cone algorithm, but not with a significant ΔROC.

Second, the cone algorithm uses all selected genes, while the ANN uses only a few principal components. Thus, more information is used in the cone algorithm. To look into how important this is for the performance of the cone algorithm, we

have tried to use principal components in the cone algorithm instead of genes. We found that good results were maintained using the 4 largest components, while the performance got noticeable worse when using only the 3 largest components.

8.3 Combination of gene expressions and clinical observables

Classifiers based on a combination of genes and clinical observables did not outperform the ones based on clinical observables alone.

When combining the two kinds of data, the 7 clinical observables and the two major principal components of the gene expression data were used as inputs to the ANNs. The two major principal components were the best selection of inputs when using expression data only. When combining data, the optimal approach would rather be to find the principal components most complementary to the clinical observables. Unfortunately, such an optimization would probably be specific to the training samples, since the committee of ANNs that used ranking of principal components in the gene expression ANN classifier almost never had the best generalization performance on the validation samples.

Hence, we cannot conclude that the information on metastatic risk available in the gene expression levels is independent of conventional clinical observables. However, we cannot conclude the opposite without more training samples that would decrease the risk of overtraining when choosing the principal components most complementary to the clinical observables.

The cone algorithm and the ANN classifier using clinical observables have many of their misclassified samples in common. All of the metastasis positive samples misclassified by the cone algorithm is also misclassified by the ANN using clinical observables and the 0.5 cutoff, and all of the metastasis negative samples misclassified by this ANN classifier is also misclassified by the cone algorithm.

8.4 Prediction of ER positive samples

When we looked closer at the result based on clinical observables for 97 samples, we noted that many misclassified samples were ER negative. Ignoring the ER negatives in the output list gave a much larger ROC area, 0.84. To confirm this result, we redid the whole analysis on the 68 ER positive samples. The resulting ROC areas, presented in table 1, are all in very good agreement with the ROC areas obtained by excluding the ER negative samples from the corresponding output list based on 97 samples. This indicates that the ER negative samples are of little help in predicting the metastatic risk of the ER positive samples. This is interesting, since it is well known that ER status defines tumours with distinctly different cell properties, see

e.g. [6]. Due to the few ER negative samples available in this data set (29), a more systematic investigation of the importance of ER status in metastasis prognosis must be postponed to future work.

Acknowledgements

First of all, I want to thank my supervisor Patrik Edén for support, encouragement and valuable insights. Also, I want to thank Carsten Peterson for being a driving force behind this project, Mattias Ohlsson for the help regarding ANN and ΔROC permutation tests, and Hongyue Dai for providing information essential for the understanding and reproduction of the prediction method in [1].

References

[1] van't Veer, L. J. *et al.*, *Nature* **415**, 530-536 (2002)

[2] van de Vijver, M. J. *et al.*, *N Engl J Med* **347**, 1999-2009 (2002)

[3] Software available at the Department of Theoretical Physics, Lund University.

[4] Quackenbush, J., *nature genetics supplement* **32**, 496-501 (2002)

[5] Supplementary information Roberts, C. J. *et al.*, *Science* **287**, 873-880 (2000)

[6] Gruvberger, S. *et al.*, *Cancer Res.* **61**, 5979-5984 (2001)