A neural network model for estrogen receptor status prediction

Iulian Nastac
Åbo Akademi University, TUCS/IAMSR,
Lemminkäisenkatu 14B, FIN-20520 Turku, Finland
e-mail: Iulian.Nastac@abo.fi

Yrjö Collan
University of Turku, Department of Pathology,
Kiinamyllynkatu 10, FIN-20520 Turku, Finland
e-mail: Yrjo.Collan@tyks.fi

Barbro Back
Åbo Akademi University, IAMSR,
Lemminkäisenkatu 14B, FIN-20520 Turku, Finland
e-mail: bback@abo.fi

Mikael Collan
Åbo Akademi University, IAMSR,
Lemminkäisenkatu 14B, FIN-20520 Turku, Finland
e-mail: mcollan@abo.fi

Päivi Jalava
University of Turku, Department of Pathology and Department of Internal Medicine, Kiinamyllynkatu 10, FIN-20520 Turku
e-mail: paivi.jalava@tyks.fi

Teijo Kuopio
Department of Pathology, Hospital of Central Finland, Keskussairaalantie 19, FIN-40620 Jyväskylä, Finland
e-mail: teijo.kuopio@ksshp.fi

T U C S

Turku Centre for Computer Science
TUCS Technical Report No 610
April 2004
ISBN 952-12-1351-5
ISSN 1239-1891

Abstract

This paper reports results on using an artificial neural network (ANN) for predicting the estrogen receptor (ER) status, which is not always available, but has a place in therapy selection of breast cancer. Our results show that in more than two thirds of the cases, the ANN is able to predict the correct ER status. An optimum neural architecture was researched, and optimal cutpoint for prediction was selected on the basis of clinical data.

Keywords: ER, neural network, training, test, prediction, cutpoint, efficiency, sensitivity, specificity.

Data Mining and Knowledge Management Laboratory

1. Introduction

Artificial Neural Networks (ANNs) are modelling tools having the ability to adapt to and to learn complex topologies of inter-correlated multidimensional data. The goal of our research was to find a mathematical model describing the relationship between 4 (or 3) inputs and one-output variables. The medical problem in this paper was as follows: Can estrogen receptor (ER) status of breast cancer be predicted with the help of clinical data, and with what level of accuracy? ER status has a place in therapy selection, but the status is not always available. The tumor may be so small (during this era of breast cancer screening) that there is not tissue enough for determining the ER status. In certain locations (e.g. in developing countries or in unusually warm weather conditions) immunohistochemical status cannot be determined because of the extra expenses it causes, or because the necessary laboratory facilities cannot be created under conditions in which the interruptions of electric main supply and the weather make the use of refrigerators unreliable.

Earlier studies clearly indicate that cancers showing high concentration of ERs have characteristic morphological features [1]: cells and nuclei are smaller and more uniform than in cancers with low concentration or absence of ERs. Correspondingly it is well known that ER-positive tumors are better differentiated, have lower histological grade and better prognosis [2]. ER determination plays a role in deciding about the use of antiestrogenic therapy [3]. Harbeck et al. [4] forecasted the relapse-free survival by using ER among other different parameters as input of an ANN model. At the time when the determination of the ER status was based on a biochemical test, Baak and Persijn [5] tested whether ER positivity could be predicted by morphometric measurements, and found it possible.

Today, the ER receptor status is evaluated by immunohistochemistry, and we decided to try the ANN in determination of the ER status when only basic clinical features are available.

The structure of this paper is as follows. Section 2 presents the problems that concern medical data, model structure and training procedure. The main features of our experimental results, regarding efficiency, sensitivity and specificity are given in next section, where we also discuss specific aspects. Conclusions are formulated in the last section of the paper.

2. Description of the data

There were two sets of data under study, which had 487 (Set I) and 387 (Set II) rows of patients. Each set of data included: age of the patient, tumor size (in cms), nodal status (presence or absence of cancer foci in the regional lymph nodes), histological grade of the neoplasm (from G1 to G3) and ER (percent of ER positive nuclei). The first set (487 samples) was from Jyväskylä Central Hospital, Jyväskylä, Finland. The estrogen receptor (ER) status was based on ER staining of the imprint upon a cut surface of breast cancer tissue. The positive staining was evaluated with an image analytic methodology on a glass slide by the CAS 200 instrument (Cell Analysis Systems, Inc., Division of Becton Dickinson, Inc., Elmhurst, IL 60126-4944, USA).

The positivity was determined as the fraction of stained nuclei of all nuclei in the sample.

The second set (387 samples) was from Turku University Hospital, Finland. The method for evaluating ER positivity was based on ER staining on paraffin sections of breast cancer. The area in the paraffin section, which showed the most consistent staining was evaluated, and in that area the fraction of positive nuclei of all nuclei was defined as ER positivity. The method used for determining ER positivity slightly differed in the two sets of data, but the form of presentation was the same in the two datasets, and the positivity was evaluated by the fraction of positive nuclei of all nuclei.

Table 1. Minimum and maximum values in the two datasets studied (note the comparable range of datasets values).

	Age		**Size**		**Nodal status**		**GR**		**ER**	
	min	**max**	**min**	**max**	**min**	**max**	**min**	**max**	**min**	**max**
Set I (training)	26	95	0.2	10	0	1	1	3	0	97.4
Set II (test)	31	98	0.17	15	0	1	1	3	0	100

Since there is no reason to believe that the two methodological approaches to evaluate ER positivity could give different results, the datasets were considered comparable in terms of ER staining methodology.

3. Model structure and training procedure

A good choice of the training data set is not a trivial task when one wants to make a good prediction. Data preprocessing and data selection remain essential steps in the knowledge discovery process for real world applications and greatly improve the network's ability to capture valuable information when correctly carried out [6]. In Fig. 1 we present our idea in training a feedforward ANN for predicting ER positivity on the basis of age of the patient, tumor size, nodal status, and histological grade.

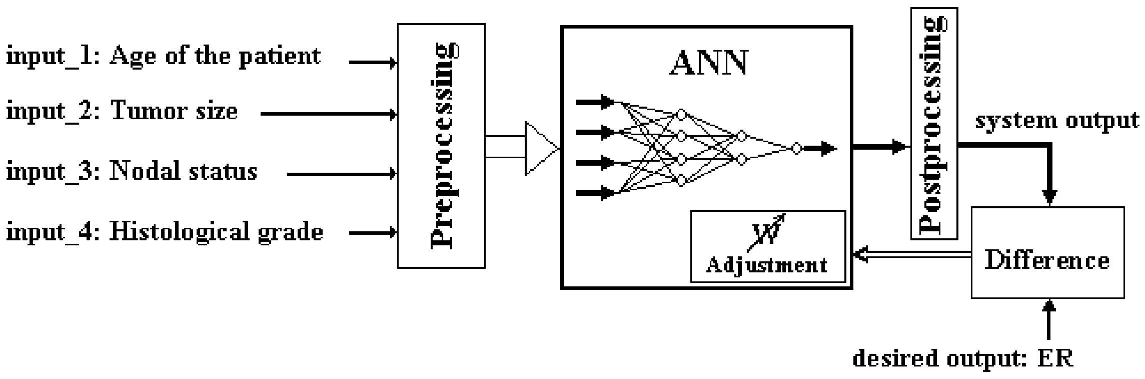

Figure 1. Training process of the feedforward ANN.

We preprocessed the inputs and outputs using normalization. We applied the reverse process of normalization in defining the simulated outputs. During the training process the ANN used the correlations that existed between the age, size, nodal status and grade, on one hand, and ER on the other. The basic training algorithm was the Scale Conjugate Gradient (SCG) algorithm [7]. In order to avoid the over-fitting phenomenon we have applied the early stopping method [8] during the training process. As splitting criterion we have randomly chosen approximately 85% of the Set I for training and the remaining for validation.

We designed a feedforward ANN with two hidden layers in order to achieve a good approximation function because in our preliminary research we obtained better results for two hidden layers than for one hidden layer, maintaining a similar ratio (approx. 10/1) between the number of the training samples and the total number of the weights. First step was performed in order to decide the proper number of neurons for each hidden layer (N_{h1} and N_{h2}). Each of the trainings started with the weights initialized to small uniformly distributed values. We chose the best model according to the smallest error between the desired and simulated outputs. This error (E_{tr}) was calculated for data that include both training sets. The supplementary condition for the error of validation sets (E_{val}) was:

$$E_{val} \leq \frac{6}{5} \cdot E_{tr}$$

We tested several architectures with different combinations of N_{h1} and N_{h2}, where:
$2 \leq N_{h1} \leq 6$ and $2 \leq N_{h2} \leq N_{h1}$.

4. Experimental results

After the training process with Set I our tool predicted the ER status of Set II. In order to evaluate the results, we used a cutpoint that assigns the original immunohostochemical values of ER versus ANN outputs in a 2×2 table (Fig. 2a).

The numbers of samples in each square allow the evaluation of the efficiency, sensitivity and specificity (as well as the fractions of false positives and false negatives) of the ANN method in determining the ER status [9]:

$$Efficiency = \frac{N_I + N_{IV}}{N_I + N_{II} + N_{III} + N_{IV}} \cdot 100$$

$$Sensitivity = \frac{N_I}{N_I + N_{II} + q} \cdot 100$$

$$Specificity = \frac{N_{IV}}{N_{III} + N_{IV} + q} \cdot 100$$

where N_k is the number of points in square k, and q is a small constant value (q=0.001) used to avoid improper computations (like 0/0).

By varying the value of cutpoint between 0 and 100 we have obtained the following graphs that show the evolutions of the predictions for Set I, and Set II. We split the problem in two cases where we used ANN with four (Case 1), and three (Case 2) inputs, respectively.

In Case 1 (*four inputs: Age, Size, Node, Grade*), after an iterative process, we tested 25 ANN architectures for each combination of N_{h1} and N_{h2}. We chose the best model according to the smallest error between the desired and simulated outputs after training process. Consequently we obtained: N_{h1}= 5 and N_{h2}= 4.

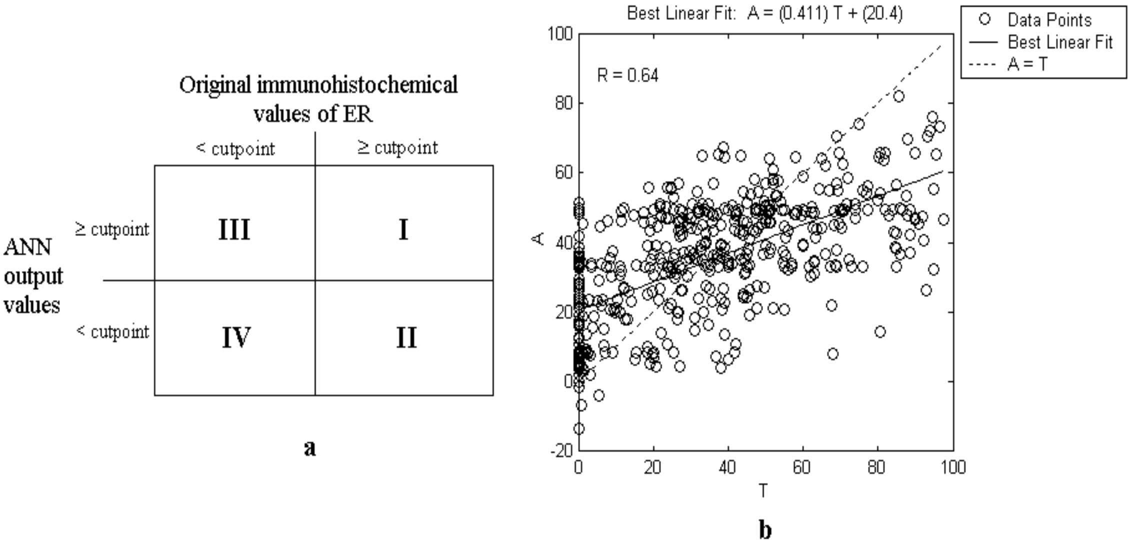

Figure 2. The squares of the correspondences between ANN outputs and originally defined immunohistochemical results (**a**), and ANN values (A) vs. target values (T) for Set I (**b**).

Plotting the ANN outputs versus original values indicated the accuracy of results (Fig. 2b). The correlation coefficient (R-value) between the outputs (A) and targets (T) was 0.64 and the regression line did not show the expected (45 degree) inclination. The significance of this result revealed us that the value of ER could not be perfectly approximated by using age of the patient, tumor size, nodal status, and histological grade. For instance, in the data sets there were some patients who presented similar pattern of the input parameters but different values of ER. Moreover we found (Fig. 2b) a few negative values of the ANN output that were provided by validation set, which was used during the training process as an indicator for stopping the training. It was not complicated to automatically adjust all negative values to zeros but we preferred to show the real values of the ANN output. The imperfect correlation was the reason why we used the cutpoint to estimate the status of the system output (positive or negative). The results were basically similar with both available datasets.

We also wanted to test at which scale cutpoint the ER value gave the best performance with respect to specificity, sensitivity, or efficiency.

The fraction of squares I + IV (Fig. 2a) is equal to the efficiency of the method in correctly predicting individual cases (Fig. 3, Table 2). For Set I (training set), efficiency varies between 70% and 100% (Fig. 3a). Values were lowest at the cutpoints 25-55, highest at cutpoints 1-20 and cutpoints greater than 60. The lowest efficiency (about 70%) is reached in the vicinity of cutpoint 35, and it is in line with the falsely evaluated patients (about 30% in the same region). The maximum fraction of false negatives (around 15%) is seen at 50.

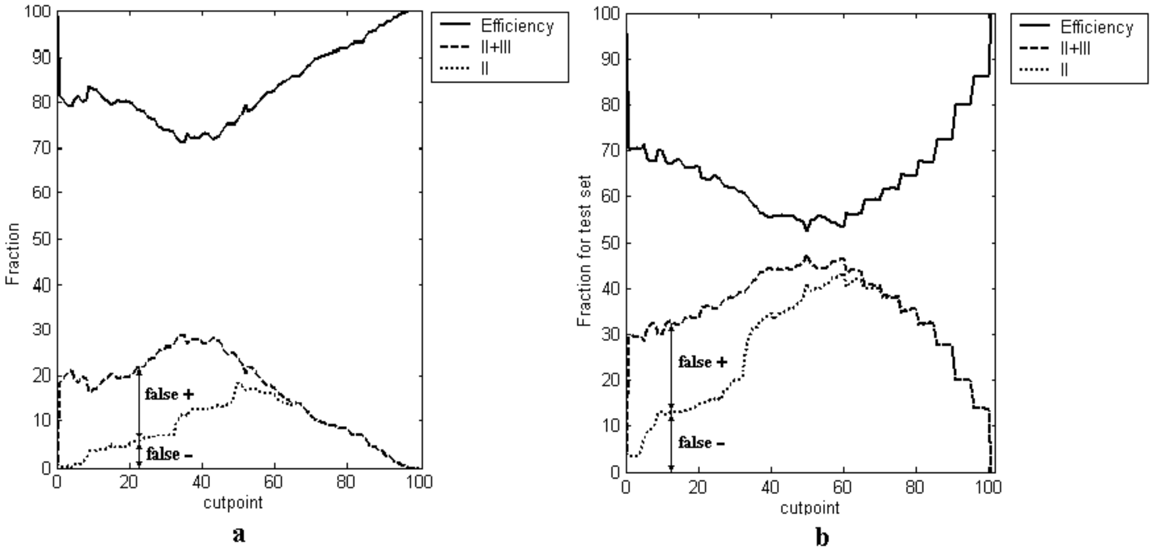

Figure 3. Efficiency, fraction of false results (from squares II+III), and fraction of false negative results (square II) versus cutpoint for Set I (**a**), and for Set II (**b**), respectively. The fraction of false positive results (sq. III) is the interval between the two graphs at the bottom of the figures.

In Fig. 4a we show the sensitivity and the specificity for Set I. Sensitivity decreases abruptly after the cutpoint value 30. There is a remarkable improvement of specificity above the cutpoint value of 20.

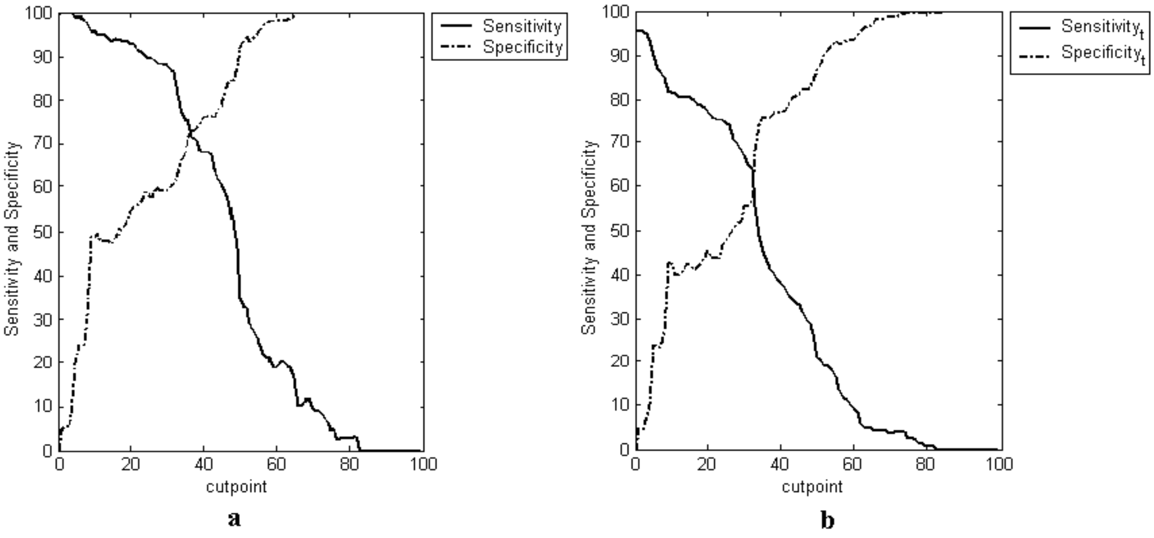

Figure 4. Sensitivity and specificity versus cutpoint for Set I (**a**), and for Set II (**b**), respectively.

Corresponding data (sensitivity, specificity, efficiency, and fractions of false positives and false negatives) for Set II (test set) are shown in Fig. 3b and Fig. 4b. We noticed the same behavior as the one presented for Set I (training set). It is quite clear for the Set II that efficiency is lower (minimum 55%, in Fig. 3b, versus 70% in Fig. 3a), and also sensitivity and specificity are lower than in the training set (Set I). For sensitivity and specificity, the average difference is about 10% over the whole range of the scale, but there are places where the difference is up to 30 percent units (Fig. 4a versus Fig. 4b, in the vicinity of cutpoint 40). Consequently, the fractions of false results (false negatives, especially) are higher in the Set II (Fig. 3a versus Fig. 3b). All these phenomena are due to the fact that Set II was not used during the training process of ANN.

The parameter age might not have the same relevance in Northern Europe as in the developing countries. Therefore we repeated the experiment without age as input. In Case 2 (*three inputs: Size, Node, Grade*), the best model, according to the smallest error between the desired and simulated outputs, had $N_{h1} = 5$ and $N_{h2} = 3$. We noticed that in this case the graphs are similar with those that were already presented in Fig. 3 and Fig. 4.

Table 2 shows efficiency, sensitivity, and specificity values, respectively, for ANNs with age input and without age input at eleven values of the cutpoint. These values are selected in the interval [10, 60] that is considered reasonable for medical purposes. The results for the training set (Set I) and the test set (Set II) are shown separately. It seems that the cutpoint 20, which is usually used in clinical practice, performs well and gives reasonably high sensitivity and efficiency figures. However, at this cutpoint only half of the negative cases are truly negative (specificity around 50%). Change in the cutpoint to improve specificity will necessarily lead to lower sensitivity and efficiency values. At the cutpoint 20 at least two thirds of the situations will be correctly evaluated by ANN (lowest efficiency 66.4%).

Table 2. Efficiency, sensitivity, and specificity versus cutpoint for Case 1: with age as input (Set I / Set II) and Case 2: without age as input (Set I / Set II).

	Efficiency				***Sensitivity***				***Specificity***			
	With age as input		**Without age as input**		**With age as input**		**Without age as input**		**With age as input**		**Without age as input**	
Cutpoint value	**Set I**	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**	**Set I**	**Set II**
10	82.9569	70.0258	75.77	72.3514	95.5056	81.6176	98.5955	95.2206	48.8549	42.6087	13.7404	18.2609
15	79.4661	67.9587	78.2341	72.093	93.5294	80.6949	93.5294	87.6448	46.9387	42.1875	42.8571	40.625
20	**79.8768**	**66.4083**	**77.8234**	**68.9922**	**92.8571**	**77.4703**	**92.8571**	**75.8893**	**54.5454**	**45.5224**	**48.4848**	**55.9701**
25	76.7967	64.3411	72.8953	65.1163	89.6193	74.074	92.3875	69.9588	58.0808	47.9166	44.4444	56.9444
30	73.922	62.2739	69.8152	60.9819	87.5	66.8103	80.0781	56.4655	58.8744	55.4838	58.4415	67.7419
35	71.2526	58.3979	70.2259	59.6899	75.1111	45.045	76.000	53.1531	67.9389	76.3636	65.2672	68.4848
40	72.8953	55.5556	70.4312	52.7132	68.3937	38.2488	61.1399	34.1014	75.8503	77.647	76.5306	76.4705
45	72.6899	55.814	70.8419	50.646	60.606	33.1707	53.9394	19.5122	78.882	81.3186	79.5031	85.7142
50	75.77	52.4548	73.7166	48.8372	34.5323	20.8955	17.2662	12.4378	92.2414	86.5591	96.2643	88.172
55	79.2608	55.5556	78.4394	53.7468	25.2252	16.3158	9.9099	7.89473	95.2127	93.401	98.6702	97.9695
60	82.5462	53.23	81.1088	53.7468	18.9473	9.18918	5.26315	5.4054	97.9592	93.5643	99.4898	98.0198

It is worth to mention that the age input does not significantly improve efficiency, sensitivity or specificity. The average improvements ($\bar{I}_p$) with age (in percent units), presented in Table 3, are computed by using the following formula:

$$\bar{I}_p = \frac{\sum_{k=1}^{N} I_p(k)}{N}$$

where: $I_p(k) = p_{Case1}(k) - p_{Case2}(k)$ is the improvement for cutpoint k, and N = 100 is the total number of cutpoints considered.

Table 3. Average improvements ($\bar{I}_p$) and improvements for cutpoint 20.

	Efficiency		***Sensitivity***		***Specificity***	
	Set I (training + validation set)	**Set II** (test set)	**Set I** (training + validation set)	**Set II** (test set)	**Set I** (training + validation set)	**Set II** (test set)
$\bar{I}_p$	1.44559	$-5.04 \cdot 10^{-15}$	3.94526	1.18234	2.16132	-0.888556
$I_p(20)$	2.05339	-2.58398	$1.84 \cdot 10^{-13}$	1.58103	6.0606	-10.4478

We noticed that there are slightly better results (positive values) in Case 1 with respect to efficiency, sensitivity and specificity for training set (Set I) and also sensitivity for test set (Set II). But, there are slightly better results in Case 2 with respect to efficiency and specificity just for test set (Set II). Specificity and efficiency present opposite tendencies as regards Set I and Set II, for Case 1 in contrast with Case 2. This effect is appreciable for cutpoint 20 with respect to specificity. Therefore we appreciate that the average improvements with age are relatively small and not consistent.

5. Discussion and conclusions

The prognostic value of various clinical features is variable. The best prognosticators are tumor size, lymph node status or mitotic activity [10]. Even though the mitotic activity is an extremely efficient prognosticator [11] we did not consider it in our study, because the feature is not always available. On the other hand, tumor size and lymph node status can, and are consistently evaluated, for every patient who has gone through the surgical treatment. We decided to consider histological grade as a variable in our ANN model because the grade is determined for every patient and also includes subjective evaluation of mitotic activity. Since ER-positive breast cancers are better differentiated and they have better prognosis, one can expect that the three prognostic features could predict ER positivity.

Our method exploits the correlations that exist among previous mentioned parameters. During the training process, ANN tried to reach the values defined by immunohistochemichal methodology. Perfect regression is, however, impossible to reach.

Training can be improved if we use more data and eventually more input parameters. Adding a new parameter as input must be done carefully, since the age, for example, might not play an important role in predicting the ER status.

Current research targets the implementation of an adaptive system, which will be periodically retrained, in order to continuously improve the model accuracy using new medical databases.

6. References

[1] A. Guazzi, C. Bozzetti, M. I. Riva, D. Zaffe and G. Cocconi, “Relationship between estrogen receptor concentration and cytomorphometry in breast cancer”, Cancer, Vol. 56, 1985, pp. 1972-1976.

[2] P. Jalava, Y.U.I. Collan, T. Kuopio, L. Juntti-Patinen and P. Kronqvist, “Bcl-2 immunostaining: A way to finding unresponsive postmenopausal N+ breast cancer patients”, Anticancer Res., Vol. 20, 2000, pp. 1213-1220.

[3] R.B. Dickson and M.E. Lippman, “Cancer of the breast. Molecular biology of breast cancer”, Chapter 37, Section 1, in Cancer. Principles and practice in oncology, DeVita et al. (Eds.), 6th Edition, Lippincott Williams & Wilkins, Philadelphia, 2001, pp. 1633-1651.

[4] N. Harbeck, R. Kates, K. Ulm, H. Graeff and M. Schmitt, “Neuran network analysis of follow-up data in primary breast cancer”, Int. J. Biol. Markers, Vol. 15, 2000, pp. 116-122.

[5] J.P.A. Baak and J.P. Persijn, “In search for the best qualitative microscopical or morphometrical predictor of oestrogen receptor in breast cancer”, Pathol. Res. Pract., Vol. 178, 1984, pp. 307-314.

[6] J. S. Armstrong, Principles of Forecasting: A Handbook for Researchers and Practitioners, Kluwer Academic Publishers, Boston, 2001.

[7] M. F. Moller, “A scaled conjugate gradient algorithm for fast supervised learning”, Neural Networks, Vol. 6, 1993, pp. 525-533.

[8] H. Demuth and M. Beale, Neural Network Toolbox, The MathWorks, Inc., Natick, 2001.

[9] R. S. Galen and S. R. Gambino, Beyond Normality: The Predictive Value and Efficiency of Medical Diagnoses, John Willey and Sons, New York, 1975.

[10] J. P. A. Baak, H. van Dop, P. H. J. Kurver and J. Hermans, “The value of morphometry to classic prognosticators in breast cancer”, Cancer, Vol. 56, 1985, pp. 374-382.

[11] P. Kronqvist, T. Kuopio, Y. Collan, “Morphometric grading in breast cancer: thresholds for mitotic counts”. Hum. Pathol., Vol. 29, 1998, pp. 1462-1468.

Turku Centre for Computer Science
Lemminkäisenkatu 14
FIN-20520 Turku
Finland

http://www.tucs.fi/

University of Turku
- Department of Information Technology
- Department of Mathematics

Åbo Akademi University
- Department of Computer Science
- Institute for Advanced Management Systems Research

Turku School of Economics and Business Administration
- Institute of Information Systems Science