DRAFT VERSION of paper to appear at the Oncology Reports, special issue Computational Analysis and Decision Support Systems in Oncology, last quarter 2005

The Wisconsin Breast Cancer Problem: Diagnosis and DFS time prognosis using probabilistic and generalised regression neural classifiers

Ioannis Anagnostopoulos[1]
Christos Anagnostopoulos[2]
Angelos Rouskas[1]
George Kormentzas[1]
Dimitrios Vergados[1]

[1]Department of Information and Communication Systems Engineering,
University of the Aegean, Karlovassi 83200, Samos – GREECE

[2]Department of Cultural Technology and Communication
University of the Aegean, Mytiline 81100, Lesvos – GREECE

contact authors: janag@aegean.gr

Abstract

This papers deals with the breast cancer diagnosis and prognosis problem employing two proposed neural network architectures over the Wisconsin Diagnostic and Prognostic Breast Cancer (WDBC/WPBC) datasets. A probabilistic approach is dedicated to solve the diagnosis problem, detecting malignancy among instances derived from the Fine Needle Aspirate (FNA) test, while the second architecture estimates the time interval that possibly contain the right end-point of the patient's Disease-Free Survival (DFS) time. The accuracy of the neural classifiers reaches nearly 98% for the diagnosis and 92% for the prognosis problem. Furthermore, the prognostic recurrence predictions were further evaluated using survival analysis through the Kaplan-Meier approximation method and compared with other techniques from the literature.

1. Introduction

According to the American National Cancer Institute, the population of the estimated new breast cancer cases for the 2004 in U.S., was near 215000, while the estimation of deaths was more than 40000 (excluding the in citu cases) [1]. In addition, the National Cancer Institute of U.S. estimates that 13.4 percent of women born today will be diagnosed with breast cancer at some time in their lives [2]. For the diagnosis of the breast cancer cases as well as for the prognosis of the disease many techniques have been discussed [3], [4], [5], [6], [7], [8], [9] and [10]. The method that can confirm malignancy with high-level sensitivity is the surgical biopsy yet is considered as a costly operation, which has a negative impact over the patient's psychology. Towards these considerations, machine leaning techniques target to provide the same levels of accuracy, without the negative aspects of surgical biopsy.

This paper deals with the breast cancer diagnosis and prognosis problem using the Wisconsin Diagnostic Breast Cancer (WDBC) as well as the Wisconsin Prognostic Breast Cancer (WPBC) data sets, which are publicly available by anonymous ftp [11]. These data sets involve measurements taken according the Fine Needle Aspirate (FNA) test. The role of diagnosis is to provide a distinction between the malignant and benign breast masses. In case that a patient is diagnosed with breast cancer, the malignant mass must be excised. After this or a different post-operative procedure, a prediction of the expected course of the disease must be determined. However, prognostic prediction does not belong either on the classic learning paradigms of function approximation or classification. This is due to a patient can be classified as a "recur" case (instance) if the disease is observed, while there is no a threshold point at which the patient can be considered as a "non-recur" case. The data are therefore censored since a time to recur for only a subset of patients is known. For the others patients, the length of time after treatment during which malignant masses are not found is known. This time interval is the disease free survival (DFS) time, which can be reported for an individual patient or for a study population. In particular, the right endpoints of the recurrence time intervals are right censored, as some patients will inevitably change hospital, doctors or die of other unrelated with the cancer causes. Therefore, the training dataset for the learning phase is not well-defined. Several groups have approached prognosis as a separation problem using different learning architectures such as backpropagation artificial neural networks [12], entropy maximization networks [13], [14] decision trees [15] and fuzzy-based measurements [16].

2. Materials and Methods

The WDBC and WPBC datasets are the results of the efforts made at the University of Wisconsin Hospital for the diagnosis and prognosis of breast tumours solely based on FNA test. This test involves fluid extraction from a breast mass using a small-gauge needle and then visual inspection of the fluid under a microscope. Figure 1 depicts two images, which were taken from fine needle biopsies of breast tumours [15].

Two neural network architectures are proposed in this paper for the breast cancer detection/prognosis problems. The first is a probabilistic classifier, which can detect malignancy while the second architecture consist of a probabilistic neural network that employs a generalised regression algorithm, estimating the recurrence time (TTR, time-to-recur) as well as a period where the patient exceeds her disease-free survival (DFS) time. The prognosis of the specific time interval is considered a difficult problem since the training data are right censored [12], [13], [17] and [18].

2. 1 Data sets

The Wisconsin Diagnostic Breast Cancer (WDBC) dataset consist of 569 instances (357 benign – 212 malignant), where each one represents FNA test measurements for one diagnosis case. For this dataset each instance has 32 attributes, where the first two attributes correspond to a unique identification number and the diagnosis status (benign / malignant). The rest 30 features are computations for ten real-valued features, along with their mean, standard error and the mean of the three largest values ("worst" value) for each cell nucleus respectively. These ten real values, which are depicted at Table 1, are computed from a digitized image of a fine needle aspirate (FNA) of breast tumour, describing characteristics of the cell nuclei present in the image and are recorded with four significant digits.

The Wisconsin Prognostic Breast Cancer (WPBC) dataset consists of 198 instances (151 non-recur - 47 recur), where each one represents follow-up data for one breast cancer case. These were consecutive in-patients at the University of Wisconsin Hospital, the period from 1984 to 1995 and include only those cases exhibiting invasive breast cancer and no evidence of distant metastases at the time of diagnosis. Each instance has 35 attributes, where the first three attributes correspond to a unique identification number and to the prognosis status (recur / non-recur) following by the recurrence time (Time To Recur - TTR) or the DFS time respectively. Then they follow the above-mentioned 30 features, while the last two attributes are the diameter of the excised tumour (in centimetres) and the number of positive axillary lymph nodes observed at time of surgery.

For the addressed problem, both WDBC and WPBC datasets were used in several publications in the medical literature [10], [15], [19], [20], [21], [22]. In addition, due to their consistency and robust creation, these datasets are also used for verification purposes over the classification or prediction performance of information systems in other scientific areas [23], [24].

2.2 Neural Networks

2.2.1 DiagnosisNN (PNN)

For the diagnosis problem the neural network involved belongs to the probabilistic type (PNN), since this kind of networks present high-generalization ability and do not require large amount of training data [25]. In the problem addressed, the PNN decides whether the input instance corresponds to benign or malignant case.

The diagnosis is based over a classification procedure, which is stated as sampling an s-component multivariate random vector $X=[x_1, x_2, ..., x_s]$, where the samples are indexed by u, $u=1,...,U$ [26]. Knowing the probability density functions for all vector populations, classification decisions are consequently made in accordance to Equation 1, which defines the Bayes optimal decision rule, where h_k stands for the probability that a sample will be drawn from population k and c_k stands for the cost of misclassifying the sample.

$d_m(X) = h_m c_m f_m(X)$,

if $h_m c_m f_m(X) > h_n c_n f_n(X)$ for all populations where m≠n **(1)**

The topology of the PNN was 31-568-2. The input layer consists of 31 nodes, which correspond to the diagnosis status, followed by the 30 calculated values (mean, standard error, "worst" value) from the digitised image of each instance. Every node receives the respective feature value through a vector, which is called Patient Record (PR). The second layer is the middle/pattern layer, which organises the training set in such a way, that an individual processing element represents each normalised input vector. Therefore, it consists of 568 nodes, which correspond to the total amount of the patterns for the training epoch, which will be described in the followings. Finally, the network has an output layer consisting of 2 nodes, representing the decision upon malignancy or not.

A conscience full competitive learning mechanism is between the weights of the input and middle layer and tracks how often the outputs win the competition with a view of equilibrating the winnings, implementing an additional level of competition among the elements to determine which processing element is going to be updated. Assuming that *O* is the number of outputs, the weight update function for the winning output is defined in Equation 2. In this equation y_i is the ith output vector that measures the distance between the input and the output neurons' weight vectors, x_j is the jth input vector and iw_{ij} is the connection weight that links the processing element j with the processing element i. Finally, f_i corresponds to the output's frequency of winning where $0 \leq f_i \leq O^{-1}$ and b_i defines the respective bias vector created by the conscience mechanism.

$$U(y) = max_i(y_i + b_i) = max_i((\sum_{j=1}^{31}(x_j - iw_{ij})^2)^{\frac{1}{2}} + b_i), \qquad i = 1, 2, \ldots, 193 \qquad \textbf{(2)}$$

where: $b_i = \gamma \cdot [O \cdot (\beta \cdot (1 - f_i))]$, i : winner

$b_i = \gamma \cdot [O \cdot (\beta \cdot f_i)]$ $\quad i$: otherwise

$\beta=0.01, \gamma=0.3, O=568$

Between layers, the activation of a synapse is given by the Euclidean distance metric and the function, which mimics the neuron's synaptic interconnections is defined by $m(x_j(t), iw_{ij}) = (\sum_j (x_j(t) - iw_{ij})^2)^{\frac{1}{2}}$. The cost function is given by $J(t) = \frac{1}{2}\sum_j (d_j(t) - m(x_j(t), iw_{ij}))^2$, where d_j is every desired response during the training epoch. Towards the optimisation of the cost function, $\frac{\partial J(t)}{\partial m(x_j(t), iw_{ij})} = 0$ should be satisfied.

In parallel, the proposed PNN uses a supervised training set to develop distribution functions within the middle layer. These functions are used to estimate the likelihood of the input instance being part of a learned WDBC instance dataset. The middle layer represents a neural implementation of a Bayes classifier, where the time interval class dependent probability density functions are approximated, using the Parzen estimator. This estimator is generally expressed by $\frac{1}{n\sigma}\sum_{i=0}^{n-1} W\left(\frac{x - x_i}{\sigma}\right)$, where n is the sample size, χ and χ_i are the input and sample points, σ is the scaling parameter that controls the area's width considering the influence of the respective distance and *W* is the weighting function. This approach provides an optimum pattern classifier in terms of minimising the expected risk of misclassifying an object. With the estimator, the approach gets closer to the true underlying class density functions, as the number of training samples increases, provided that the training set is an adequate representation of the class distinctions. The likelihood of an unknown instance belonging to a given class is calculated according to Equation 3.

$$g_{i(PR)} = \frac{1}{(2\pi)^{\frac{p}{2}} \sigma^p N_i} \sum_{j=0}^{(N_i - 1)} e^{\frac{-(PR - \bar{x}_{ij})^T (PR - \bar{x}_{ij})}{2\sigma^2}} \qquad \textbf{(3)}$$

In the above equation, *i* reflects to the number of the class, *j* is the pattern layer unit, $\bar{x}_{ij}$ corresponds to the jth training vector from class *i* and *PR* is the Patient Record. In addition, N_i represents the respective training vectors for the i^{th} class, *p* equals to the *PR* dimension (*p*=31), σ is the standard deviation and $(2\sigma)^{-2}$ outlines the beta (β) coefficient. In other words, Equation 3 defines the summation of multivariate spherical Gaussian functions centred at each of the training vectors $\bar{x}_{ij}$ for the i^{th} class probability density function estimate.

Furthermore, in the middle layer, there is a processing element for each input vector of the training set and equal amounts of processing elements for each output class, in order to avoid one or more classes being skewed incorrectly. Each processing element in this layer is trained once to generate a high output value when an input *PR* matches the training vector (training instance). However, the training vectors do not need to have any special order in the training set, since the category of a particular vector is specified by the desired output. The learning function simply selects the first untrained processing element in the correct output class and modifies its weights to match the training vector. The middle layer operates competitively and only the highest match to an input *PR* prevails and generates an output.

2.2.2 PrognosisNN (GRNN)

For the prognosis problem the system involved is a Generalised Regression Neural Network architecture (GRNNs). These neural networks have the special ability to deal with sparse and non-stationary data where non-linear relationships exist among inputs and outputs. In the problem addressed, the network calculates a time interval that corresponds to a possible right end-point of the patient's disease-free survival time.

Thus, if $f(x,z)$ is the probability density function of the vector random variable x and its scalar random variable z, then the GRNN calculates the conditional mean $E(z \setminus x)$ of the output vector. The joint probability density function (pdf) $f(x,z)$ is required to compute the above conditional mean. GRNN approximates the pdf from the training vectors using Parzen windows estimation, which is a non-parametric technique approximating a function by constructing it out of many simple parametric probability density functions. Parzen windows are considered as Gaussian functions with a constant diagonal covariance matrix according to Equation 4.

$$E(z \setminus x) = \frac{\int_{-\infty}^{\infty} z \cdot f(x,z)dz}{\int_{-\infty}^{\infty} f(x,z)dz} \text{ , where} \qquad \textbf{(4)}$$

$$f_p(x \setminus z) = \frac{1}{(2\pi\sigma^2)^{(\mathrm{N}+1)/2}} \cdot \frac{1}{P}\sum_{i=1}^{P}\left(e^{\frac{-D_i^2}{2\sigma^2}} \cdot e^{\frac{-(z-x_i)^2}{2\sigma^2}} \right)$$

In the above equation *P* equals to the sample points x_i, *N* is the dimension of the vector of sample points, D_i is the Euclidean distance between *x* and x_i calculated by $D_i = \|x - x_i\| = \sqrt{\sum_{i=1}^{N}(x-x_i)^2}$, where *N* is the amount of the input units to the network. Additionally, σ is a width parameter, which satisfies the asymptotic behaviour as the number of Parzen windows becomes large according to Equation 5 where *S* is the scale. When an estimated pdf is included in $E(z \setminus x)$, the substitution process defines Equation 6 in order to compute each component z_j.

$$\lim_{P\to\infty}(P\sigma^N(P)) = \infty \text{ and}$$

$$\lim_{P\to\infty}\sigma(P) = 0 \text{, when } \sigma = \frac{S}{P^{E/N}} \text{ , } 0 \le E < 1 \qquad \textbf{(5)}$$

$$z_j(x) = \frac{\sum_{i=1}^{P} z_j^i c_i}{\sum_{i=1}^{P} c_i} \text{, where } c_i = e^{\frac{-D_i^2}{2\sigma^2}} \qquad \textbf{(6)}$$

However, due to the fact that the computation of Parzen estimation is a time consuming procedure when the sample is large, a clustering procedure is often incorporated in GRNN, where for any given sample x_i, instead of computing the new Gaussian kernel k_i at centre *x* each time, the distance of that sample to the closest centre of a previously established kernel is found, and the old closest kernel *(k-1)*$_i$ is used again. Taking into account this approach, Equation 6 is transformed to Equation 7.

$$z_j(x) = \frac{\sum_{i=1}^{P} A_i(k)c_i}{\sum_{i=1}^{P} B_i(k)c_i} \text{ , } 1 \le j \le M \text{ , } A_i(k) = A_i(k-1) + z_j \text{ , } B_i(k) = B_i(k-1) + 1 \qquad \textbf{(7)}$$

From the input to the pattern layer a training/test vector *X* is distributed, while the connection weights from the input layer to the k^{th} unit in the pattern layer store the centre X_i of the k^{th} Gaussian kernel. In the pattern layer the summation function for the k^{th} pattern unit computes the Euclidean distance D_k

between the input vector and the stored centre X_i and transforms it through the previously described exponential function c_i. Afterwards, the *B* coefficients are set as weights values to the remaining units in the next layer. In the summation/division layer the summation function computes the denominator of Equation 6 for the first unit *j* and then the numerator for the next unit *(j+1)*. Thus, in order to compute the output of the respective function, the summation of the numerator is divided by the summation of denominator and such output is forwarded to the output layer. Finally, the output layer receives inputs from the summation/division layer and outputs are estimated conditional means, computing the error on the basis of the desired output.

The topology of the prognosis neural network was 14-193-4-4. The input layer consists of 14 nodes, which correspond to the prognosis status, the TTR or the DFS time, the ten cell nuclei characteristics attributes of Table 1, the diameter of the excised tumour and the number of positive axillary lymph nodes observed at time of surgery. It must be noted that due to the small amount of WPBC instances, the standard error and the "worst" values from the ten real-valued features were removed in order to avoid the "curse of dimensionality" problem during the training phase of the artificial neural network. Four instances were not included in the training/testing set since the Lymph node values were missing. Thus, the second layer consists of 193 nodes, which correspond to the total amount of the patterns for the training epoch. Finally, it follows the summation/division layer, which consists of 4 nodes that feed a same amount of processing elements in the output layer representing the classified time interval that correspond to a possible right-end of the DFS time (during the first year, 1-3 years, 3-6 years, more than 6 years).

The GRNN was implemented in C++ and trained in a Pentium IV, 3.2 GHz, 1024 MB RAM. The time needed for the completion of the training epoch, was approximately 2.4 minutes. During the training period, the 'beta' coefficient for all the local approximators of the pattern layer, which was used to smooth the data being monitored, was set to 100 $(\beta = 1/2\sigma^2=100)$. In addition, the mean and the variance values of the randomised biases were equal to 0 and 0.5 respectively.

3. Results

3.1 Diagnosis NN

The training set of the proposed diagnosis neural network consists of the WDBC dataset instances and its role is to classify an instance as benign or malignant. The quality of prediction was examined using the jackknife test in which, each instance was singled out in turn, as a test instance with the remaining instances used to train the neural network.

The mean time needed for the completion of one training epoch, was 4.4 seconds. Equation 8 and Equation 9, outline the Akaike's Information Criterion (AIC) as well as the Rissanen's Minimum Description Length (MDL) during the training period. Values $|d_{ij} - y_{ij}|$ correspond to the distances among the desired and the actual network output for the ith exemplar residing at the jth processing element.

$$AIC(k) = N * ln(MSE) + 2 * K \qquad \textbf{(8)}$$

$$MDL(k) = N * ln(MSE) + 0.5 * K * ln(N) \qquad \textbf{(9)}$$

Mean Square Error: $MSE = \dfrac{\sum_{j=1}^{P}\sum_{i=1}^{N}(d_{ij} - y_{ij})^2}{N \cdot P}$

In the above criteria, *P* equals to the number of the output processing elements (*P=2*), while *N* and *K* define the amount of the exemplars in the training set and the number of network weights respectively (*N=568, K=18744*). AIC measures the trade-off between training performance and network size, while MDL combines the error of the model with the number of degrees of freedom for determining the level of generalization. The afore-mentioned indicators are used in order to fine-tune the biases mean and variance of the respective local approximators and produce a confusion matrix with the best possible values in the diagonal cells. Thus, the condition that must be fulfilled is the stabilisation of AIC and MDL over the *β* coefficient. It was evaluated that the neural network was not properly trained when $0.1<\beta<1$, due to large MSE and therefore large AIC and MDL values. Conversely, MSE was significantly decreased when *β* was set to one and the training confusion matrix revealed acceptable percentage values in the diagonal cells. However, further investigation over the influence of *β* in the learning ability of the classifier, exposed that MSE and AIC/MDL values were increased for values of *β* that range between 1 and 3 and then further reduced when $3<\beta<100$. Finally, for $\beta\geq100$ the

AIC/MDL criteria were optimised (*AIC=19615* and *MDL=41621*), satisfying in parallel the condition $\partial MDL / \partial \beta = 0$. All the afore-mentioned measurements are depicted in Figure 2. Therefore, during the training period, the 'beta' coefficient for all the local approximators of the middle layer was set equal to 100 ($\beta = 100$), while the mean and the variance values of the randomised biases were equal to 0 and 0.5.

Table 2 presents the confusion matrix between the tested classes. This matrix is defined by labelling the desired classification on the rows and the predicted classifications on the columns. The diagonal cells correspond to the correctly classified web pages for each class respectively, while the other cells show the misclassified pages. Therefore, the confusion matrix displays values where each one corresponds to the percentage effect that a particular input has on a particular output.

Thus, having an "a priori" known set of 357 benign and 212 malignant instances the neural network successfully identified 348 and 209 instances respectively. This implies that the precision of the system in respect to the two WDBC categories was 97.47% and 98,58%, while the recall was 99.14% and 96.76% respectively. On the other hand, 7 benign and 3 malignant instances were erroneously misclassified as depicted in Table 2. The overall performance of the diagnosis neural network derived from the total amount of the correctly classified web pages residing in the diagonal cells, versus the total amount of the sample set (*overall performance = 557/569 ≈98%*).

3.2 Prognosis NN

The training set of the prognosis neural network consists of the WPBC instances, which were divided over four classes, namely C_1, C_2, C_3 and C_4, according to the value of the third attribute, which indicates the recurrence or the disease-free survival (DFS) time. Thus, C_1 corresponds to the instances, in which the DFS time or the recurrence time was between 1 and 12 months, while C_2, C_3 and C_4 correspond to intervals between 1-3 years, 3-6 years and more than 6 years. Table 3 depicts the amount of the WPBC dataset instances in respect to the above-mentioned categorisation. The first column indicates the time interval class, while the second and the third columns present the amount of instances, when the tumour recurred (N_R) and the amount of instances when the tumour did not recur (N_N).

The total WPBC instances were presented to the network in a round-robin manner leave-one out, 10-fold cross-validation, while the training ended before the average testing error on the left-out cases began to increase. The total prediction accuracy *(TP)* and the prediction accuracy *(P)* for each location calculated for assessment of the prediction system are given by Equations 10 and 11.

$$TP = \frac{\sum_{k=1}^{4} p_k}{N} \quad \textbf{(10)}$$

$$P = \frac{p_k}{n_k} \quad \textbf{(11)}$$

In the above equations, N is the total number of sequences (N_R+N_N), k is the respective class, n_k is the number of instances in class k and p_k is the number of correctly predicted instances in class k.

The accuracy of prediction by leave-one out tests for the WPBC instances and the categorized time intervals are shown in the confusion matrix of Table 4. The diagonal cells correspond to the correctly classified WPBC instances for each class respectively, while the other cells present the misclassified instances. Every row expresses the system's ability in terms of correct classification over a tested time interval, while the overall performance was *92.3%.*

4. Discussion

4.1 Diagnosis NN - Comparison with other techniques

In this section the results derived from the proposed work are compared with similar techniques that target to solve the Wisconsin Breast Cancer problem. These techniques are coming from the area of neural networks [27], [28], [29], support vector machines and decision trees [30], fuzzy logic-based approaches [31], [32] and Ant Colony Optimization (ACO) algorithms [33]. Among them, the proposed probabilistic neural architecture with the β selection over the AIC/MDL criteria presented the second better precision value over the WDBC dataset (precision=97.9%). A hybrid approach of neural network with fuzzy logic-based rules (Feature Space Mapping) presented a slightly better accuracy (precision=98.3%) [29], [32]. The third better performance (97.2%) was achieved by a Support Vector Machine approach, where the authors used 5-fold cross validation over the WDBC dataset [30]. Table

5 depicts a survey over proposed methods that tackle the WDBC problem. It is interesting to note that neural networks generally presented a better performance in respect to all the other approaches. Especially, this was observed when other methods (decision trees, fuzzy logic-based rules and Bayesian classification techniques), were used as stand-alone systems and not as supplementary classification modules to neural networks. Two algorithms, which follow the Ant Colony Optimization (Ant_Miner1 and Ant_Miner2) presented mean accuracy rates significantly lower than these of the neural network architectures [32]. Finally, the Quadratic Discriminant Analysis presented very poor accuracy [28].

4.2 Prognosis NN - Comparison with other techniques (in WPBC dataset)

As far as concern the prognosis problem, the recurrence predictions made by the Prognosis NN were further examined using survival analysis. The cases were divided, according to the predicted disease-free survival (DFS) time, into the aforementioned time intervals (less than a year, between 1st-3rd years, between 3rd-6th years, beyond the 6th years). The actual recurrence probabilities of these four time intervals were then assembled, using the Kaplan-Meier approximation. Figure 3 presents the survival analysis of the predicted DFS curve produced by the probabilistic neural network, compared to the actual DFS curve, over the WPBC dataset instances. The y-axis corresponds to the probability of DFS time and the x-axis corresponds to the time in months.

The two curves are almost identical for time intervals between 0-11 months, 20-28 months, 40-58 months and 80-88 months and very similar to the rest intervals except from the predictions made for the period of more than 90 months. Thus, despite the fact that the Prognosis NN had better performance for the classes that correspond to larger DFS time (Table 4), the survival analysis of the predicted results shown no significant statistical differences for lower DFS times. This caused by the “unfair” division of the learning set over the categorized intervals (Table 3), as well as due to the fact that the range of the predicted interval elongates for larger DFS times.

In a similar approach the predicted probability DFS values were always in higher levels in respect to the actual data, without important statistical differences [34]. However, this is not desirable in prognosis problems since the end of the disease-free survival time may correspond to a possible recurrence of the disease. Thus, in the proposed approach the predicted DFS time curve verifies that the Prognosis NN deals more adequately with the Wisconsin breast cancer prognosis, since in most intervals the predicted probabilities for the right-end-point of the patient’s disease-free survival time is in lower levels in respect to the actual ones.

Additionally, if the predicted right-end points of the Prognosis NN are considered as possible TTR points then the blue, green and red curves at Figure 4, depict the Kaplan-Meier disease-free survival time probabilities based on the predicted TTR that belong to three time intervals (predicted TTR ≤ 2 years, 2 years < predicted TTR ≤ 5 years and TTR ≥ 5 years). The selection of these intervals made in order to compare the results with the work described in [35], where the recurrence rate was nearly 30% at two years and only 10% recurrence up to five years, but with more observed recurrences after this interval. Conversely, the respective recurrence rates outlined from the work proposed in this paper (blue and the red curve at Figure 4) shows that the instances for which the prediction was less than two years have a better prognosis that reaches nearly 50% at two years, while those instances with predicted TTR of greater than five years have nearly 10% (same prognosis levels with the work described in [35]).

5. References

1. http://seer.cancer.gov/cgi-bin/csr/1975_2001/search.pl#results, Estimated New Cancer Cases and Deaths for 2004
2. U.S. National Institutes of Health, National Cancer Institute, http://cancernet.nci.nih.gov/
3. Wang, T.C., Karayiannis N.B., Detection of microcalcifications in digital mammograms using wavelets, IEEE Transactions on Medical Imaging, Vol. 17, Issue. 4, Aug. 1998, pp. 498 – 509.
4. Huo, Z., Giger, M., Vyborny, C., Wolverton, D., Schmidt, R., Doi, K., Automated computerized classification of malignant and benign mass lesions on digital mammograms, Acad. Radiol. 5, 155–168, 1998.
5. Cheng Heng-Da, Lui Yui Man, Freimanis R.I., IEEE Transactions on Medical Imaging, Vol. 17, Issue. 3, June 1998, pp. 442 – 450.
6. Pendharkar P.C., Rodger J.A., Yaverbaum G.J., Herman N. and Benner M., Association, statistical, mathematical and neural approaches for mining breast cancer patterns, Expert Systems with Applications, 17:223–232, 1999.

7. Setiono R., Generating concise and accurate classification rules for breast cancer diagnosis, Artificial Intelligence in Medicine, 18:205–219, 2000.
8. Chen D., Chang R.F., Huang Y.L., Breast cancer diagnosis using self-organizing map for sonography, Ultrasound in Medical Biology 2000, Vol. 26, pp. 405–11.
9. Giger M., Huo Z., Kupinski M., Vyborny C., Computer-aided diagnosis in mammography. In Handbook of Medical Imaging, (Eds.) Sonka, M., Fitzpatrick, J., Medical Image Processing and Analysis, Vol. 2. SPIE Press, pp. 917–986, 2000.
10. Tourassi G.D., Markey M.K., Lo J.Y., Floyd Jr. C.E., A neural network approach to breast cancer diagnosis as a constraint satisfaction problem, Med. Phys. Vol.28, pp. 804–811, 2001.
11. http://ftp.ics.uci.edu/pub/machine-learning-databases/breast-cancer-wisconsin/, Wisconsin Diagnostic Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset.
12. Burke H. B., Goodman P.H., et al, Artificial neural networks improve the accuracy of cancer survival prediction, Cancer, Vol. 79, pp. 857-862, 1997.
13. Choong P.L, deSilva C.J.S et al., Entropy maximization networks, An application to breast cancer prognosis, IEEE Transactions on Neural Networks, 1996, 7(3):568-577.
14. Choong P.L., deSilva C.J.S, Maximum entropy estimation vs. multivariate logistic regression: which should be used for the analysis of small binary outcome data sets?, Proceedings of the 20th Annual International Conference of the IEEE Engineering in Medicine and Biology Society, vol.3, pp:1602 – 1605, 1998.
15. Wolberg W.H., Street W.N., Heisey D.M., and Mangasarian O.L., Computer-derived nuclear features distinguish malignant from benign breast cytology, Human Pathology, 26:792--796, 1995.
16. Seker H., Odetayo M., Petrovic D., Naguib R.N.G., Bartoli C., Alasio L., Lakshmi M.S., Sherbet G.V., A fuzzy measurement-based assessment of breast cancer prognostic markers, Proceedings of the 2000 IEEE EMBS International Conference on Information Technology Applications in Biomedicine, 9-10 Nov. 2000, pp.174 – 178.
17. Mangasarian et al, “Breast cancer diagnosis and prognosis via linear programming”, Operations Research, 43(4), pp. 570-577, July-August 1995.
18. Street W. N., “A neural network model for prognostic prediction”, Proceedings of theFifteenth International Conference on Machine Learning, Madison, Wisconsin, Morgan Kaufmann, 1998.
19. Wolberg W.H., Street W.N., and Mangasarian O.L., Machine learning techniques to diagnose breast cancer from fine-needle aspirates, Cancer Letters 77 (1994) 163-171.
20. Wolberg W.H., Street W.N., and Mangasarian O.L., Image analysis and machine learning applied to breast cancer diagnosis and prognosis, Analytical and Quantitative Cytology and Histology, Vol. 17, No. 2, pages 77-87, April 1995.
21. Jiang Y., Nishikawa R., Wolverton D., Metz C., Giger M.L., Schmidt R., Doi K., Automated feature analysis and classification of malignant and benign microcalcifications, Radiology 198, 671–678, 1996.
22. Taylor P., Fox J. and A. Todd-Pokropek, Evaluation of a decision aid for the classification of microcalcifications, Digital Mammography, Nijmegen: Kluwer Academic Publishers, pp. 237-244, 1998.
23. Hoya T. and Chambers J. A., “Heuristic pattern correction scheme using adaptively trained generalized regression neural networks”, IEEE Trans. Neural Networks, vol.12, no.1, pp. 91-100, 2001.
24. Kaban A., Girolami M., Initialized and guided EM-clustering of sparse binary data with application to text based documents, 15th International Conference on Pattern Recognition, Vol.2 pp.744-747, Sept. 2000.
25. Specht D.F., ‘Probabilistic Neural Networks’, Neural Networks, vol.3, no.1, pp.109-118, 1990.
26. Masters T., ‘Advanced Algorithms for Neural Networks’, John Wiley: New York, 1995.
27. LTF-C: Architecture, Training Algorithm and Applications of New Neural Classifier, Marcin Wojnarski, Fundamenta Informaticae, pp. 89-105, Volume 54, No 1, 2003
28. B. Ster and A. Dobnikar, *Neural networks in medical diagnosis: Comparison with other methods*. In A. Bulsari *et al*., editor, Proceedings of the International Conference EANN '96, pages 427-430, 1996.
29. New developments in the Feature Space Mapping model, Rafal Adamczak Wlodzislaw Duch, Czκstochowa, Third Conference on Neural Networks and Their Applications, October 14-18, Poland, 1997.
30. K.P. Bennett, J. Blue, A Support Vector Machine Approach to Decision Trees, R.P.I Math Report No. 97-100, Rensselaer Polytechnic Institute, Troy, NY, 1997

31. H.J. Hamilton, N. Shan, N. Cercone, RIAC: a rule induction algorithm based on approximate classification, Tech. Rep. CS 96-06, Regina University 1996.
32. W. Duch, R. Adamczak & K. Grabczewski, "A new methodology of extraction, optimization and application of crisp and fuzzy logical rules", IEEE Trans Neural Networks, 11(2), 1-31, 2000.
33. Classification Rule Discovery with Ant Colony Optimization, Liu B., Abbass HA and McKay B, IEEE Computational Intelligence Bulletin, Vol.3, No.1, pp.31-35, February 2004.
34. A Neural Network Model for Prognostic Prediction, Street WN, Proceedings of the 15th International Conference on Machine Learning, pp.540-546, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA , 1998.
35. W. N. Street, O. L. Mangasarian and W.H. Wolberg, An inductive learning approach to prognostic prediction, 12th International Conference on Machine Learning, pp. 522-530, A. Prieditis and S. Russell (eds), Morgan Kaufmann, San Francisco, 1995.

List of Tables

cell nuclei characteristics
1. radius [mean of distances from centre to points on the perimeter],
2. texture [standard deviation of grey-scale values],
3. perimeter,
4. area,
5. smoothness [local variation in radius lengths],
6. compactness [((perimeter)^2 / area) - 1],
7. concavity [severity of concave portions of the contour],
8. concave points [number of concave portions of the contour],
9. symmetry,
10. fractal dimension [“coastline approximation” – 1]

Table 1. WDBC/WPBC cell nuclei characteristics attributes

Diagnosis NN		**Predicted**		
		B	**M**	**Precision (%)**
Actual	**B**	348	7	97,47
	M	3	209	98,58
Recall (%)		99,14	96,76	**97,89**

Table 2. Testing phase Confusion Matrix (B:benign, M:malignant – Diagnosis NN)

Class	Interval time	N_R	N_N
C_1	Less than 1 year	20	23
C_2	1 year – 3 years	14	34
C_3	3 years – 6 years	7	48
C_4	More than 6 years	5	43

Table 3. WPBC instances according to the categorised interval time and prognosis status

		Predicted				
		C_1	C_2	C_3	C_4	**Precision (%)**
Actual	C_1	39	2	1	1	90,69
	C_2	1	44	1	2	91,67
	C_3	1	1	51	2	92,72
	C_4	1	1	1	45	93,75
					Total Precision	92,27

Table 4. Testing confusion matrix

Prediction method	**Precision (%)**	**Reference**
Feature Space Mapping	98.3	[29], [32]
Diagnosis NN (min. β over AIC/MDL)	97.9	This work
Support Vector Machine (5xCV)	97.2	[30]
3-NN standard Manhatan	97.1	[27]
kNN with DVDM distance	97.1	[27]
21-NN standard Euclidean	96.9	[27]
Fisher linear discr. Analysis	96.8	[28]
Multi Layer Perceptron / Back Propag.	96.7	[28]
LVQ	96.6	[28]
kNN, Euclidean/Manhattan	96.6	[28]
NB – naïve Bayes	96.4	[28]
C4.5 (decision tree)	96.0	[31]
Linear Discriminant Analysis	96.0	[28]
OC1 DT (5xCV)	95.9	[30]
GTO DT (5xCV)	95.7	[30]
Assistant I tree (ASI)	95.6	[28]
Rule Induction over Approx. Classification	95.0	[31]
Assistant R tree (ASR)	94.7	[28]
Lookahead Feature Construction binary tree	94.4	[28]
Ant_Miner3	94.3*	[33]
C4.5 (5xCV)	93.4	[30]
Ant_Miner1	92.6*	[33]
Quadratic Discriminant Analysis (QDA)	34.5	[28]

Table 5. Precision rates of Wisconsin Breast Cancer problem techniques (* mean value)

List of Figures

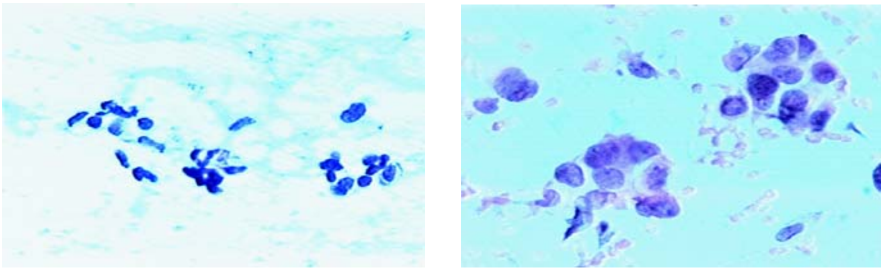

Figure 1. Images taken using the FNA test: (a) Benign, (b) Malignant

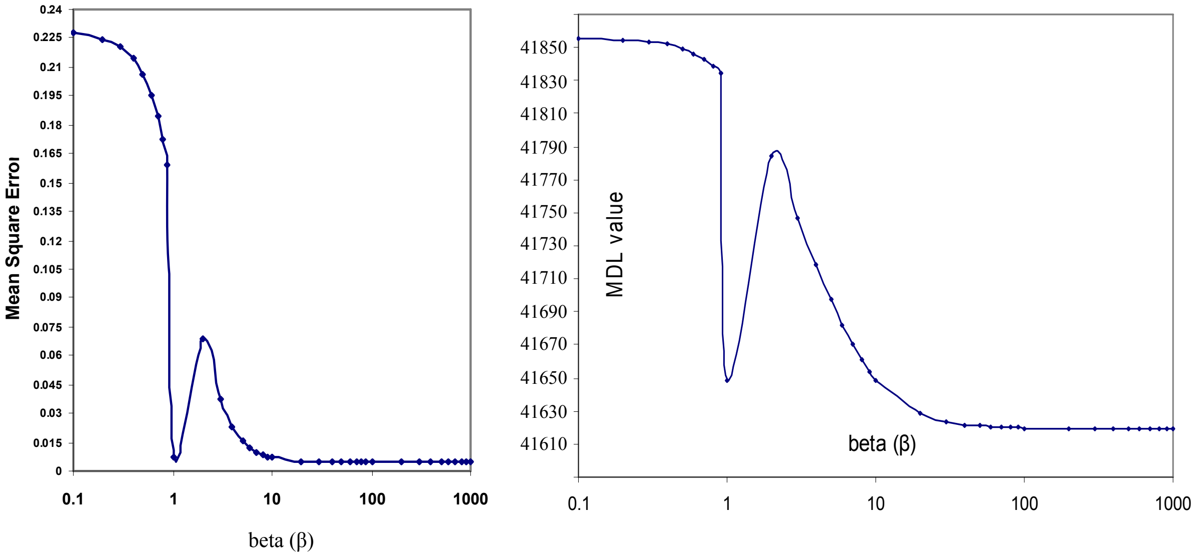

Figure 2. MSE/AIC/MDL over the beta coefficient

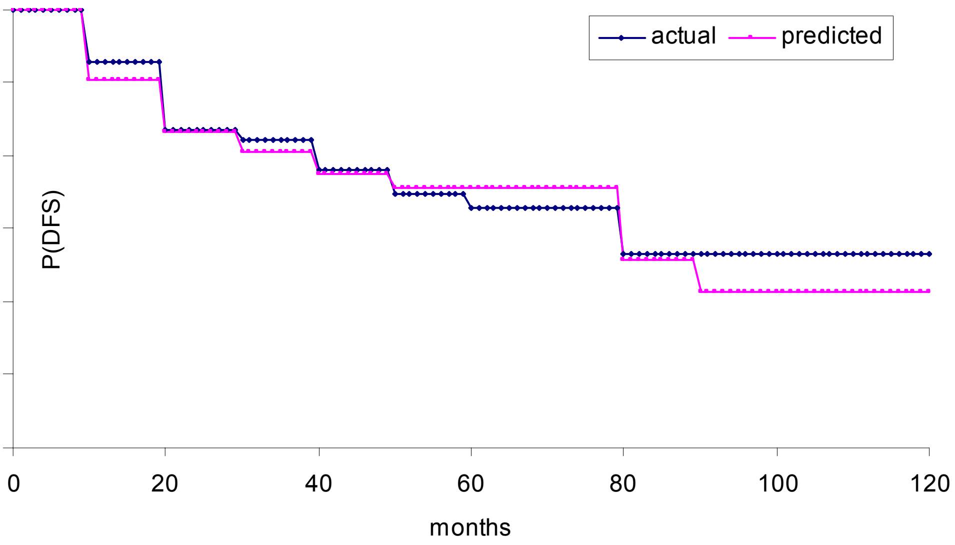

Figure 3. Survival analysis of the predicted DFS curve compared to the actual DFS curve [y-axis: DFS time probability, x-axis time in months, Dataset:WPBC]

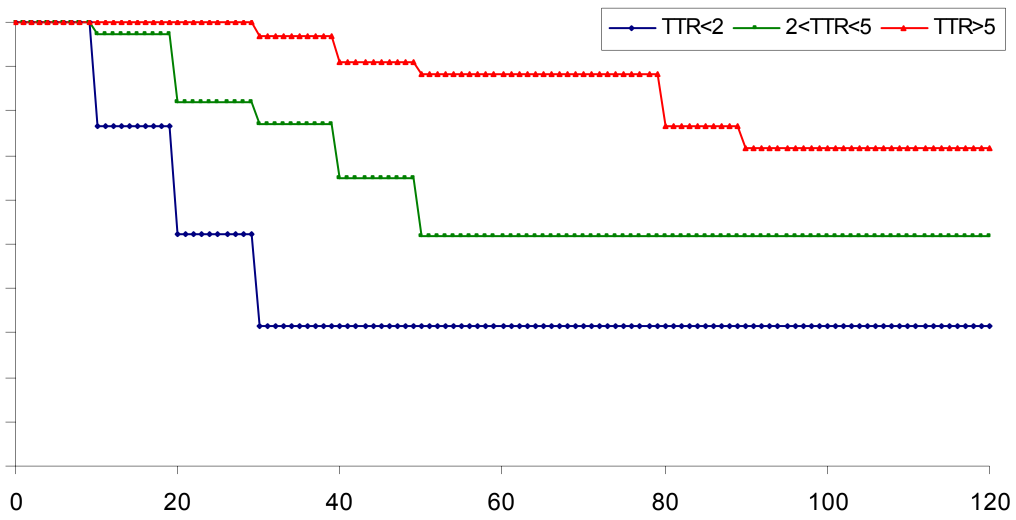

Figure 4. Kaplan-Meier disease-free survival time probabilities based on the predicted TTR (predicted TTR ≤ 2 years, 2 years < predicted TTR ≤ 5 years and TTR ≥ 5 years)