2010 2nd International Conference on Software Technology and Engineering(ICSTE)

Predicting Breast Cancer Survivability Using Data Mining Techniques

A. Soltani Sarvestani, A. A. Safavi
School of Electrical and Computer Engineering
Shiraz University
Shiraz, Iran
a.soltani.s@ieee.org,
safavi@shirazu.ac.ir

N.M. Parandeh
Department of Computer Science
University of Mysore
Mysore, India
m.n.parandeh@gmail.com

M.Salehi
Department of Pathology
Shiraz Medical Science University
Shiraz, Iran
msalehy@pearl.sums.ac.ir

***Abstract*—In this paper, appropriate and efficient networks for breast cancer knowledge discovery from clinically collected data sets are investigated. Invoking various data mining techniques, it is desired to find out the percentage of disease development, using the developed network. The results, help in choosing a reasonable treatment of the patient. Several neural network structures are evaluated for this investigation. The performance of the statistical neural network structures, self organizing map(SOM), radial basis function network (RBF), general regression neural network (GRNN) and probabilistic neural network (PNN) are tested both on the Wisconsin breast cancer data (WBCD) and on the Shiraz Namazi Hospital breast cancer data (NHBCD). To overcome the problem of high dimension of the data set and realizing the correlated nature of the data, principal component techniques are used to reduce the dimension of data and find appropriate networks. The results are quite satisfactory while presenting a comparison of effectiveness each proposed network for such problems.**

Keywords-Breast Cancer,Data Mining, Neural Network.

I. INTRODUCTION

Today, in the United States for example, approximately one in eight women over their lifetime has a risk of developing breast cancer. A major class of problems in medical science involves the diagnosis of the disease, based upon various tests performed upon the patient. When several tests are involved, the ultimate diagnosis may be difficult, even for a medical expert. This invokes, over the past few decades, to computerized diagnostic tools, intended to aid the physician in making sense out of the confusion of data [1].

There have been various studies to classify the breast cancer as were presented in [1] and [2]. However, because of incorporating more parameters in the NHBCD in comparison with WBCD investigation somehow different and more complicated. Furthermore, in the NHBCD the cancer is diagnosed and these groups of patients are going under surgery, but in WBCD the "yes" or "no" answer are only considered. In this paper, the data is used to develop various networks.

This paper provides a comparison among the capabilities of various neural networks such as Multilayer Perceptron (MLP), Self Organizing Map (SOM), Radial Basis Function (RBF) and Probabilistic Neural Network (PNN) were used to classify WBCD and NHBCD data.

The paper is followed as below: in Section 2, the problem and the databases are explained, in Section 3, data mining concepts and techniques are briefly presented. Section 4, provides a short review of some classes of neural networks that used in this works and abilities of each network are discussed. In Section5, principal component techniques and their applications are reviewed. Finally, the results of simulation are presented.

II. PROBLEM DESCRIPTION

The WBCD was obtained from the University of Wisconsin Hospitals, Madison, from Dr. William H. Wolberg(available in [9] and [14]). The database contains of 699 samples which have 683 complete data and 16 samples which have not complete attributes. There are 9 integer-valued attributes and each data value range from 1 to 10, as follows:

(1) Clump Thickness; (2) Uniformity of Cell Size;

(3) Uniformity of Cell Shape; (4) Marginal Adhesion;

(5) Single Epithelial Cell Size; (6) Bare Nuclei;

(7) Bland Chromatin; (8) Normal Nucleoli; (9) Mitoses.

These attributes measure the external appearance and internal chromosome changes in nine different scales. There are two values in the variable class of breast cancer: Benign (non-cancerous) and malignant (cancerous). The NHBCD was obtained from Namazi Hospital, Shiraz, Iran by Dr. Mohammad Salehi. The database contains of 100 samples which have 73 complete data and 27 samples which have lost attributes. There are 12 integer-valued attributes and each data has a different range:

(1) Age; (2)Grade; (3)Stage; (4) Lymph Node Group;

(5)Nottingham Prognostic Index; (6)Size Group;

(7)Estrogens Receptor; (8)Progesterone Receptor;

(9)E-Cadherin; (10)K c-erb; (11)Ki67;

(12)Follow up Duration (m).

There are two values in the variable class of breast cancer. In this case the goal was to diagnose of patient recovery the survivability. If the output is 1, the patient will successfully cure and recover. If the output is 2, the surgery will be

978-1-4244-8666-3/10/$26.00 © 2010 IEEE

unsuccessful so the patient should be given analgesic until death.

III. DATA MINING

Data mining is a process that uses a variety of data analysis tools to discover patterns and relationships in data that may be used to make valid predictions. The first and the simplest analytical step in data mining is to describe the data. Data mining is summarized its statistical attributes, review it visually using charts and graphs. Another task, look for potentially meaningful links among variables. Collecting, exploring and selecting the correct data are critically important. But, data description cannot alone provide an action plan. You must build a predictive model based on patterns which are determined from known results, and then test that model on results outside the original sample. A good model should never be confused with reality, but it can be a useful guide to understand your business. The final step is to verify the model empirically. There are two keys to success in data mining. First, is coming up with a precise which you are formulation of the problem trying to solve. A focused statement usually results in the best payoff. The second key is using the right data. After choosing some data from the available data, or buying external data, you may need to transform and combine them in significant ways.

Neural networks are of particular interest because they offer means of efficiently modeling large and complex problems in which there may be hundreds of predictor variables that have many interactions. Actual biological neural networks are incomparably more complex. Neural nets may be used in classification problems (where the output is a categorical variable) or for regressions (where the output variable is continuous) [3].

IV. NEURAL NETWORK STRUCTURE

A. Feed forward Neural Network(FNN)

Artificial neural networks are computing tools constructed of many simple interconnected elements called neurons with a unique capability of recognizing underlying relationships between input and output events. A neuron has two components: (i) a weighted sum $S = \sum w_i x_i + b$ that performs a weighted summation of the inputs (x1, x2, x3, . . . , xn) where b is the bias of the network and (ii) a linear, nonlinear or logic transfer function which gives an output corresponding to S. Here, many kinds of functions, including threshold (logic), sigmoid, hyperbolic tangent, Gaussian and linear could be used. In this paper, hyperbolic tangent (tansig) function [f(x) = 1/(1 + exp(-x)] is used in input and hidden layers. In a typical ANN, there are three types of neurons: (a) input neurons that may receive external data, (b) output neurons that send data out of the ANN, and (c) hidden neurons whose signals remain within the ANN and connect the input layer neurons to output layer neurons [4].

B. Radial Basis Functions (RBF) Network

A radial basis function (RBF) is a real-valued function whose value depends only on the distance from the origin, so that $\phi(x) = \phi(\|x\|)$; or alternatively on the distance from some other point $\boldsymbol{c}$, called a *center*, so that $\phi(x, c) = \phi(\|x - c\|)$. Any function ϕ that satisfies the property $\phi(x) = \phi(\|x\|)$ is a radial function. The norm is usually Euclidean distance. Radial basis functions are typically used to build up function approximations of the form

$$y(x) = \sum_{i=1}^{N} \omega_i \phi(\|x - c_i\|) \tag{1}$$

where the approximating function $y(\boldsymbol{x})$ is represented as a sum of N radial basis functions, each associated with a different center $\boldsymbol{c}_i$, and weighted by an appropriate coefficient w_i. Approximation schemes of this kind have been particularly used in time series prediction and control of nonlinear systems exhibiting sufficiently simple chaotic behavior.

The sum can also be interpreted as a rather simple single-layer type of artificial neural network called a radial basis function network, with the radial basis functions taking on the role of the activation functions of the network [5].

C. Probabilistic Neural Networks(PNN)

The PNN introduced by Specht is essentially based on the well-known Bayesian classifier technique commonly used in many classical pattern-recognition problems. Consider a pattern vector x with m dimensions that belongs to one of two categories K_1 and K_2. Let $F_1(x)$ and $F_2(x)$ be the probability density functions (pdf) for the classification categories K_1 and K_2, respectively. From Bayes' discriminant decision rule, x belongs to K_1 if

$$\frac{F_1(x)}{F_2(x)} \rangle \frac{L_1}{L_2}\frac{P_2}{P_1} \tag{2}$$

conversely, x belongs to K_2 if

$$\frac{F_1(x)}{F_2(x)} \langle \frac{L_1}{L_2}\frac{P_2}{P_1} \tag{3}$$

where L_1 is the loss or cost function associated with misclassifying the vector as belonging to category K_1 while it belongs to category K_2, L_2 is the loss function associated with misclassifying the vector as belonging to category K_2 while it belongs to category K_1, P_1 is the prior probability of occurrence of category K_1, and P_2 is the prior probability of occurrence of category K_2. In many situations, the loss functions and the prior probabilities can be considered equal. Hence the key to using the decision rules given by equations (2) and (3) is to estimate the probability density functions from the training patterns. In the PNN, a nonparametric

estimation technique known as Parzen windows is used to construct the class-dependent probability density functions (pdf) for each classification category required by Bayes' theory. This allows determination of the chance a given vector pattern lies within a given category. Combining this with the relative frequency of each category, the PNN selects the most likely category for the given pattern vector. Both Bayes' theory and Parzen windows are theoretically well established, have been in use for decades in many engineering applications, and are treated at length in a variety of statistical textbooks. If the *j*th training pattern for category K_1 is *xj*, then the Parzen estimate of the pdf for category K_1 is

$$F_1(x) = \frac{1}{(2\pi)^{m/2}\sigma^m n}\sum \exp\left[-\frac{(x-xj)^T(x-xj)}{2\sigma^2}\right] \quad (4)$$

where n is the number of training patterns, *m* is the input space dimension, *j* is the pattern number, and σ is an adjustable smoothing parameter. However, the choice of σ in general has been found to be not too sensitive to variations in its value [6].

D. Principal Component Analysis (PCA)

Sometimes, the dimension of the input vector is large, but the components of the vectors are highly correlated (redundant). It is useful in this situation to reduce the dimension of the input vectors by some projection methods. PCA is an effective procedure for reducing the dimensionality of large data sets. It permits identification of associations between variables, therefore reducing the dimensionality of the data set. This technique has three effects: (a) It orthogonalizes the components of the input vectors so that it can produce the uncorrelated orthogonal variables or PCs by multiplying the original correlated variables with the eigenvector. (b) It orders the resulting orthogonal components (principal components) so that those with the largest variation come first. (c) It eliminates those components that contribute the least to the variation in the data set. The input vectors should be normalized, so they have zero mean and unit variance. This is a standard procedure when we are using principal components [13]. One of the parameters in PCA method is minimum fraction variance. It determines the elimination of those principal components that contribute less than this value to the total variation in the data set. For example, if the minimum fraction variance is considered 0.02 then PCA will eliminate those principal components that contribute less than 2% to the total variation in the data set. The principal components are described by two lower dimensional data matrices (scores and loadings) that describe the underlying patterns within the original data. The scores matrix is the data formed by transforming the original data into the space of the principal components (PCs). Rows of score correspond to observations and columns to components. The eigenvalue analysis of a p×p correlation matrix produces p pairs of eigenvalues and eigenvectors [13].

Each eigenvalue/eigenvector pair describes a principal component. The eigenvalues describe the amount of variance which is explained by each principal component and the loadings are the coordinates of the eigenvector. By ordering the eigenvectors in order to descend eigenvalues (the first largest), one can create an ordered orthogonal basis with the first eigenvector which have the direction of largest variance of the data. So, we can find directions in which the data set has the most significant amount of energy. The principal component scores are then given as linear combinations of the original standardized (auto scaled) data with the loadings as the coefficients. Principal components are extracted so that the maximum amount of variance is explained in (has the largest eigenvalue associated with) the first principal component and progressively less variance is explained for each subsequent component. Consequently, the PCA technique extracts the eigenvalues of the PCs that are the measure of variance in the observations and eigenvectors (loadings or weightings) which determines the participation of the original variables in the PCs from the covariance matrix of original variables. Suppose we have a random vector population *x*, where

$$x = (x1,...,xn)^T \quad (5)$$

and the mean of that population is denoted by

$$\mu_x = E(x) \quad (6)$$

and the covariance matrix of the same data set is

$$C_x = E\left((x-\mu_x)(x-\mu_x)^T\right) \quad (7)$$

The components of *Cx*, denoted by c_{ij}, represent the covariance between the random variable components *xi* and *xj*. The component c_{ii} is the variance of the *xi* component. The variance of a component indicates the spread of the component values around its mean value. If two components *xi* and *xj* of the data are uncorrelated, their covariance is zero ($c_{ij} = c_{ji} = 0$). According to definition, the covariance matrix is always symmetric. From a symmetric matrix such as the covariance matrix, we can calculate an orthogonal basis by finding its Eigenvalues and Eigenvectors. The eigenvectors *ei* and the corresponding eigenvalues *ki* are the solutions of the equation

$$C_x e_i = \lambda_i e_i \qquad i = 1,...,n \quad (8)$$

These values can be found, for example, by finding the solutions of the characteristic equation

$$\left|C_x - \lambda I\right| = 0 \quad (9)$$

where *I* is the identity matrix which has the same order as *Cx* and the |.| denotes the determinant of the matrix[8]. If the data vector has n components, the characteristic equation becomes order n. This is easy to solve only if n is small. Solving eigenvalues and corresponding eigenvectors is a non-trivial task, and many methods exist [10]. In this paper, the Principal components analysis of the ANN input data sets were performed to use MATLAB 7.2 (R2006a).

V. RESULTS

In this paper, the fundamental aim is to achieve an effective network with appropriate algorithm for NHBCD. Since Shiraz Namazi Hospital data set (NHBCD) is not sufficiently large, firstly, all of investigations are performed on the internationally available WBCD data set and initial evaluations of various methods are obtained.

Selecting a proper size for the network is quite important. If the network's size is too small, it will not be capable of constructing a good model for the problem. On the other hand, if it is too large, then it may be "too good" for the given problem. Being " too good" means that the network represents a function which is more complex than the existing function in the training set. Consequently, the network would perform quite poor on the rest of actual data or any other unseen data, not considered in the training set. In practice, the "best" size of the network should be obtained via trial and error [7].

First, Multilayer Perceptron networks are trained on WBCD. Table 1 lists the algorithms that are tested and the acronyms used to identify them. Four different algorithms were tested for training Multilayer Perceptron networks. The network used for the WBCD is a 9-5-5-2 network with tansig neurons in all layers.

Table 2 summarizes the results of training this network with the four different algorithms. Each entrance in Table 1 represents 30 different trials, where different random of initial weights are used in each trial. In each case, the network is trained until the squared error is less than 0.01. A few runs failed to converge for some of the algorithms, so only the top 75% of the runs from each algorithm were used to obtain the statistics.

TABLE I. PERFORMANCE ALGORITHMS THAT ARE TESTED AND THE ACRONYMS WE USE TO IDENTIFY THEM

Acronym	Algorithm
LM	trainlm - Levenberg-Marquardt
RP	trainrp - Resilient Backpropagation
CGB	traincgb - Conjugate Gradient with Powell/Beale Restarts
OSS	trainoss - One-Step Secant

TABLE II. THE PERFORMANCE OF THE VARIOUS ALGORITHMS FOR THE FEEDFORWARD NN

Algorithm	MSE	Mean time(sec)
LM	0.08	10.14
RP	0.031	7.33
CGB	0.080	4.8
OSS	0.066	13.94

TABLE III. RESULT FOR COMP LEARNING

Epoch	MSE	# of Neurons	Mean time(sec)
50	0.0556	2	53.37

Next competitive learning networks are evaluated. As before, 500 of the data in WBCD were used to train in competitive learning.

The value of Kohonen learning rate and conscience learning rate were chosen 0.1 and 0.01, respectively. To create this network 2 neurons has been used. The network is trained for 50 iterations. A total of 183 samples were applied to the networks as test data; that is, 25% of the database was used to test. The result for the competitive learning is shown in Table 3. In this table mean square error is 0.0055. Fig.3 depicts post-train for this learning.

Probabilistic neural network was also applied to WBCD database to show the performance of statistical neural networks on breast cancer data. The spread value of PNN was chosen 0.35.

The PNN gives the best accurate classification with 183 correct classifications while the competitive network has the lowest accuracy with 182 correct classifications for the training set and the MLP has 180 correct classifications. Fig.4 depicts the comparison between actual targets and predictions of the PNN.

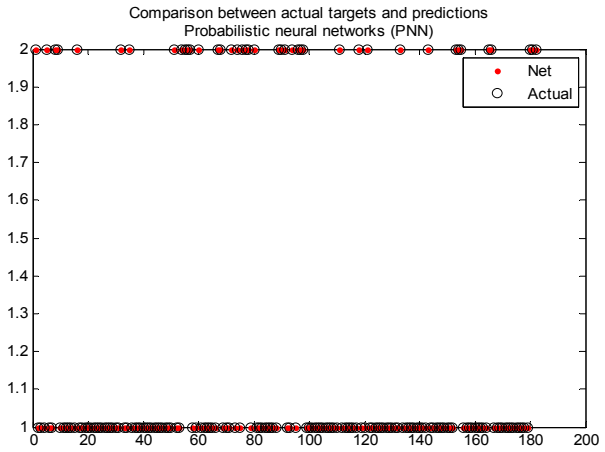

Figure 1. Probabilistic Neural Network for the WBCD.

The Shiraz Namzi Hospital data (NHBCD) is used to study. First a PNN is developed to use the first 73 training set. As the dimension of the NHBCD's input vector is large, while the components of the vectors are highly correlated, it would be a good to reduce the dimension of the input vectors. Before training, the data set is normalized to a new set that have zero mean and unit variance and after that, the principal components of the new set are extracted by considering minimum fraction variance of 0.05. It comes up with three. 60 of the data in NHBCD were used to train and 13 of the data were used to test the network. The spread value of PNN was chosen 44. Results are depicted in Table 5, and Figures 2 and 3.

TABLE IV. RESULT FOR PROBABILISTIC NEURAL NETWORK

MSE	Spread	Mean time(sec)
0.0	0.35	1.487

TABLE V. PROBABILISTIC NEURAL NETWORK FOR NHBCD USING PCA

Spread	minfrac	MSE	Train data	Test data	Mean time(sec)
44	0.05	0.00	60	13	1.57

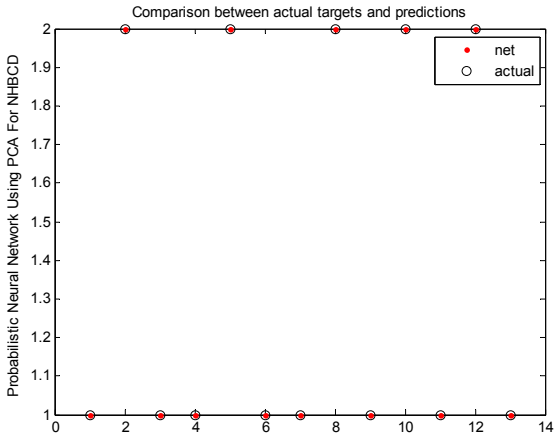

Figure 2. Probabilistic Neural Network for the NHBCD.

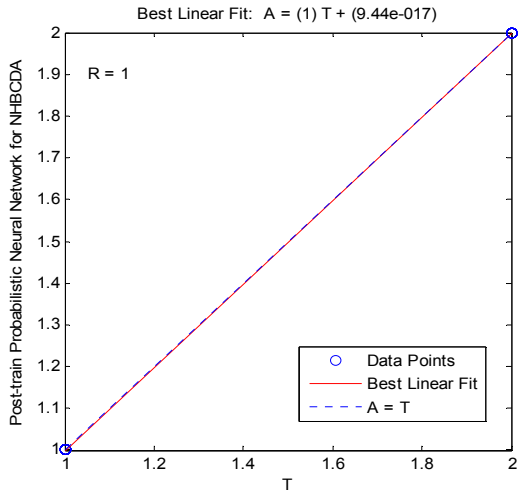

Figure 3. Post-train Probabilistic Neural Network for the NHBCD..

TABLE VI. RESULT FOR COMP LEARNING

Type	Performance	Mean time(sec)
MLP	90%	38.45
Comp	98.45%	53.37
PNN	100%	1.487

VI. CONCLUSION

This paper has investigated some studies on the use of different classes of neural networks for real clinical diagnosis of breast cancer in the largest and the most active Hospital the South Iran. The proposal was initially evaluated using some internationally available data set. By applying several neural network structures, a diagnostic system was constructed that this system performs quite well. The performance of these neural network structures was investigated for breast cancer diagnosis problem. RBF and PNN were proved as the best classifiers in the training set. However the PNN gives the best classification accuracy when the test set is considered. According to overall results, it is seen that the most suitable neural network model for classifying WBCD and NHBCD data is the PNN. This work also indicates that statistical neural networks can be effectively used for breast cancer diagnosis to help oncologists.

REFERENCES

[1] T. Kiyan, T. Yildirim, "Breast Cancer Diagnisis Using Statistical Neural Networks," JOURNAL OF ELECTRICAL & ELECTRONICS ENGINEERING,Vol.4 No.2 ,2004, pp.1149-1153.

[2] A. Bellachia, E. Guven, "Predicting Breast Cancer Survivability Using Data Mining Technigues," 2006

[3] Two Crows Corporation(2005),*Introduction to data mining and knowledge discovery*, Third edition.

[4] J. E. Dayhoff, J.E., Neural Network Architectures. Van Nostrand Reinhold, New York, 1990.

[5] S. Haykin, Neural Networks: A Comprehensive Foundation, Mac Millan College Publishing Company,1994.

[6] T. Anthony, C. Goh, "Probabilistic neural network for seismic liquefaction potential ," NRC Research Press Web site, 2002.

[7] A. A. Safavi, J.A. Romagnoli, " Application of wavelet-based neural networks to modelling and optimisation of an experimental distillation column", (IFAC Journal of) Engineering Applications of Artificial Intelligence, Vol. 10, No. 3, 1997, pp.301-313.

[8] V. Zitko, "Principal component analysis in the evaluation of environmental data," Mar. Pollut. Bull.1994, pp.718–722..

[9] www.ailab.si/orange/doc/datasets/breast-cancer-wisconsin.htm

[10] F, Benjamin et. Al "The Surveillance, Epidemiology, and End Results Program": A National Resource. Cancer Epidemiology Biomarkers & Prevention , 1999, pp.1117-1121.

[11] K. J. Cios,G. W. Moore, "Uniqueness of medical data mining." Artificial Intelligence in Medicine ,2002, pp.1-24.

[12] J. Han, M, Kamber, Data Mining concepts and Techniques, Morgan kaufmann publishers, 2001

[13] Matlab(2006). By the math works, Inc

[14] (http://www.cancer.org/).