Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 3, Issue, 7, pp.272-274, July, 2011

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

A NEURAL CLASSIFIER APPROACH FOR THE PREDICTION OF OPTIMAL BREAST CANCER ATTRIBUTES

Prof.Dr. P.K. Srimani[1] and Smt. Parvati N Angadi[2]

[1]Director, R&D (CS), Bangalore – 560 010
[2]Research Scholar, Rayalaseema University, Kurnool

ARTICLE INFO

Article History:
Received 13th February, 2011
Received in revised form
24th March, 2011
Accepted25th April, 2011
Published online 16th July 2011

Key words:
Neural Networks,
Wisconsin Breast cancer dataset,
Radial basis function,
Support vector machine, multi layer perceptron,
Accuracy, Sensitivity, Specificity.

ABSTRACT

The study pertaining to the available medical databases which are of huge size involve the consideration of all the attributes related to the problem. This resulted in large amount of computational time and accordingly there was a drastic reduction in the computational speed also. In order to minimize these factors, the present study was undertaken which enabled to get optimal results by using neural classifier techniques. In this paper, the effectiveness of various attributes and classifiers in the cytological diagnosis of WBCD breast cancer dataset were compared. Here, the most effective attributes were identified and it was found that these attributes describe at least one of the important nuclei characteristics of the morphological and textual features namely size, shape and texture of a cancerous cell. Further, applying all these attributes together to the classifiers, it was found that there was a significant increase in their performance which resulted in optimal computational speed and time. Overall, it was found that support vector machines could give accuracy as high as 97.37%.

INTRODUCTION

Breast cancer has been identified as the second largest cause of cancer related deaths among women. According to projections by the World Health Organization (WHO), the number of breast cancer diagnosis is estimated to be 1.2 million. In India, breast cancer affects one out of 22 women and in every 6.5 minutes one woman is diagnosed with the disease. Early diagnosis by means of automated breast imaging systems like digital mammography, ultrasound imaging and MRI effectively assist radiologists and physicians to give a better chance for curing from the disease and increase the survival rate (West *et al*., 2005). The breast cancer diagnosis problem has attracted many researchers in computational intelligence, data mining and statistical fields.
A good amount of research and diagnosis of the disease is found in literature. For example, Diagnosis and detection of Breast Cancer using support vector machine was made by Sudhir *et al*., 2006. Performance of statistical Neural Network structures was investigated for breast cancer diagnosis by Tüba Kıyan *et al.,* 2003. A technique based on inductive decision tree algorithm was proposed for classification problems by Aboul, 2003. A comparative analysis of different neural network techniques for binary classification problem was made by Jeatrakul *et al.,* 2009. A method for feature extraction using mutual information as criteria was proposed by O Valenzuela *et al*., 2006. Neural networks, with their remarkable ability to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques are one of the emerging and effective computational techniques. The most important neural network classifiers for real world problems are Multilayer Perceptron (MLP), Radial Basis Function (RBF) and Support Vector Machines (SVM).MLP's are feed forward neural networks trained with standard backpropogation algorithm. They are supervised networks and require a desired response to be trained. They have been shown to approximate the performance of optimal statistical classifiers in different problems (Haykin,1999).

RBF networks are non linear hybrid networks using Gaussian transfer function. The centers and widths of the Gaussians are set by unsupervised learning rules and supervised learning is applied to the output layer. Their main advantage is strong adapting ability and faster training (Haykin,1999). SVM, one of the most successful learning algorithms proposed in recent years,are set of related supervised learning methods that analyze data and recognize patterns used for classification and regression analysis. One of the main advantages of the support vector machine over other networks is it gives simple geometric interpretation. (Haykin, 1999).

***Corresponding author:**
profsrimanipk@gmail.com;parvati_angadi@yahoo.com

The main objective of the present study was to compare the effectiveness of various attributes and classifiers in the cytological diagnosis of breast cancer. Various combinations of features and classifiers were taken into consideration in order to find which performed the best in terms of accuracy, sensitivity and specificity.

MATERIALS AND METHODS

Dataset used

The Wisconsin breast cancer database (WBCD) was originally collected and provided by Dr.William.H.Wolberg and used by a number of researchers in pattern recognition and machine learning. This database was obtained from the repository of machine learning database, University of California, Irvine. It contains features that describe characteristics of the cell nuclei of a fine needle aspirate (FNA) of a breast mass (W.N Street *et al.,* 1993). For each cell nucleus, there are 30 real-valued attributes (O.L. Mangasarian, *et al.*, 1995), and 559 instances of which 357 are benign and 212 are malignant type. The three values that are provided for each attribute refer to the mean, standard error and "worst" values. Benign and malignant tissues can be differentiated by considering the difference between the visual characteristics of the nuclei of their cells. For this purpose, attributes which describe the morphological and textual characteristics are considered (W.H.Wolberg *et al.,* 1995). These attributes include: area and radius, which describe size; perimeter which expresses both size and shape; smoothness, concavity, compactness, concave points, symmetry and fractal dimension which describe shape and variance of the gray level of pixels which is an indicator of texture.

Methodology

The dataset was partitioned into two sets: training and testing. The testing set was not seen by any neural network during the training phase. It was only used for testing the generalization of neural network classifiers after they were trained. The cytological attribute (area, radius, perimeter, smoothness, concavity, compactness, concave points, symmetry and fractal dimension) values were used as input data. After many trials and experimentation, it was found that the best results were obtained by dividing 559 breast cancer samples into 455 training samples and 114 testing samples. All the three networks(multi layer perceptron, radial basis function and support vector machines) were trained by considering 455 exemplars and a maximum of 1000 epochs in supervised learning (common to all networks) while 100 epochs were used for unsupervised learning (common to radial basis function and support vector machines). The transfer function used was "TanhAxon" and the learning rule was "Levenberg Marquardt".

RESULTS AND DISCUSSION

The simulations were realized using "Neuro solutions 6.05" on WBCD database using only one attribute at a time for the classification of breast cancer into benign and malignant type. Three different Neural classifier structures, MLP, RBF and SVM were applied to make a comparative study of these and to predict an optimal performance evaluation.

Table 1. Performance of networks with the best performing attributes area, perimeter, radius and concave points

Neural Network	TP	FP	FN	TN	Sensitivity (%)	Specificity (%)	Accuracy (%)
MLP	39	2	2	71	95.12	97.26	96.49
RBF	41	4	0	69	100.00	94.52	96.49
SVM	40	2	1	71	97.56	97.26	97.37

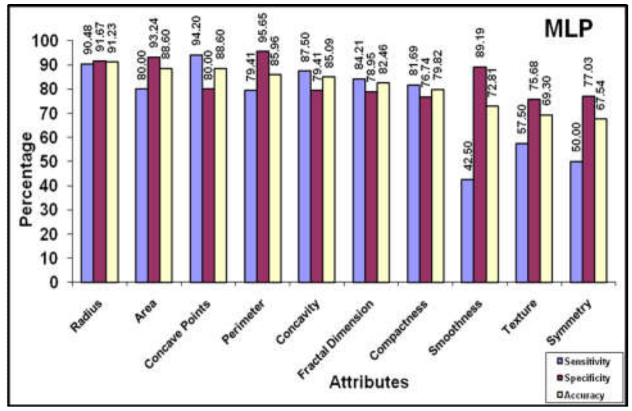

Fig 1. Performance of MLP for individual attributes

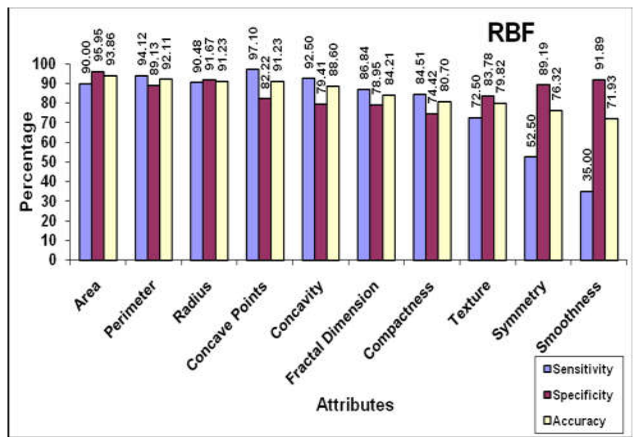

Fig 2. Performance of RBF for individual attributes

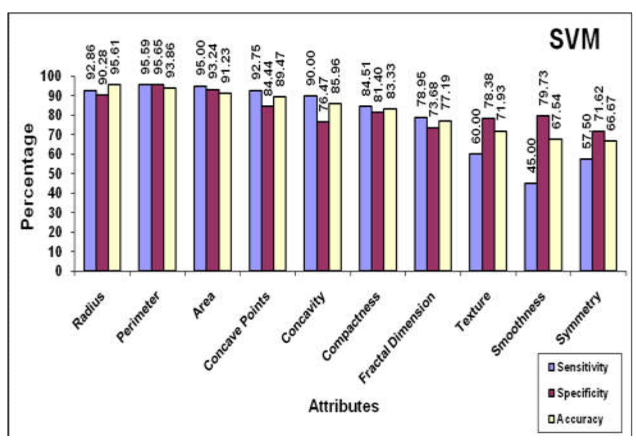

Fig. 3. Performance of SVM for individual attributes

Measures for performance evaluation

In this study, we used three performance measures namely Accuracy, Sensitivity and Specificity which can be mathematically expressed as

Accuracy= (TP+ TN)/ (TP+TN+FP+FN)
Sensitivity =TP/ (TP+FN)
Specificity=TN/ (TN+FP)
where TP, TN, FP and FN denote true positives, true negatives, false positives and false negatives respectively. A true positive decision occurs when the positive prediction of the classifier coincides with a positive prediction of the pathologist. A true negative decision occurs when both the classifier and the pathologist suggest the absence of a positive prediction. False positive occurs when the classifier labels a benign case (negative) as a malignant one (positive). Finally, false negative occurs when the system labels a positive case as negative. Further, classification accuracy is defined as the ratio of the number of correctly classified cases to the total number of cases. Sensitivity refers to the rate of correctly classified positives. Specificity refers to the rate of correctly classified negatives.

The graphs predicted that the attribute "radius" gave the highest accuracy for MLP and SVM where as the attribute "area" gave the highest accuracy for RBF. The attribute "perimeter" gave the highest specificity for MLP and SVM where as for RBF the attribute "area" gave the highest specificity. While, attribute "concave points" gave the highest sensitivity for MLP and RBF where as for SVM, the attribute "perimeter" gave the highest sensitivity. From this analysis, it was found that the attributes area, perimeter, radius and concave points contributed significantly towards greater accuracy, specificity and sensitivity. The classifiers were trained and tested considering only for these four attributes. The results were tabulated as shown in Table 1. With this best chosen set of attributes, it was found that there was a significant increase in the performance of all the three networks. By choosing this minimum set of features, irrelevant and redundant attributes were removed. Over all, SVM gave the best performance. For any further considered diagnostic samples, it would be sufficient to obtain accurate results by using samples having only these chosen attributes. With this, it was possible to optimize the computational speed and time. Figures 1,2 and 3 show the values of performance metrics for all attributes taken one at a time as the input to the MLP,RBF and SVM respectively.

Conclusions

In this paper we presented a comparative study of the effectiveness of various attributes and neural classifiers in the classification of the WBCD breast cancer dataset. The most effective attributes were identified and were applied as input to the classifiers. It was found that there was a significant increase in the performance by optimizing both speed and time. SVM yielded the most consistent results. No doubt, the results of the present investigation help the doctors to take right decisions at right time which in turn increases the survival rate.

Future scope

This paper forms the basis for further implementation and computation of different network classifiers and the other training algorithms. The work in this direction is in progress. This method can be used for the diagnosis of other diseases also.

Acknowledgement

We thank Dr. William H Wolberg at the University of Wisconsin for supporting us with the breast cancer dataset which we have used in our experiments.

REFERENCES

Aboul Ella Hassanien, March 2003. Classification and feature selection of Breast cancer data based on decision tree algorithm, *Studies in informatics and control*, Vol.12, No.1.

Jeatrakul P and Wong K.W, 2009. Comparing the Performance of Different Neural Networks for Binary Classification Problems, *Eighth International Symposium on Natural Language Processing*.

Mangasarian O.L.,Street W.N. and Wolberg W.H, 1995. Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, Vol. 43(4), pp.570-577.

Simon Haykin,1999. *Neural Networks- A Comprehensive Foundation (2nd Edition)*, Pearson India.

Street W.N, Wolberg W.H., and Mangasarian O.L., 1993. Nuclear feature extraction for breast tumor diagnosis, *IS&T/SPIE International Symposium on Electronic Imaging: Science and Technology*, Vol. 1905, San Jose, CA, 1993, pp. 861-870.

Sudhir Sawarkar D, Ashok Ghatol A., Amol Pande P., June 12-14, 2006. Neural Network Aided Breast Cancer Detection and Diagnosis Using Support Vector Machine, *Proceedings of the 7th WSEAS International Conference on Neural Networks, Cavtat, Croatia*, (pp158-163).

Tüba Kıyan, Tülay Yıldırım,2003. Breast cancer diagnosis using statistical Neural networks**,** *International XII. Turkish Symposium on Artificial Intelligence and Neural Networks – TAINN*

Valenzuela O, Rojas. I, Herrera L. J., Guillén A., Rojas F., Marquez. L.and Pasadas M, 2006. Feature selection using mutual Information and neural networks, *Monografías del Seminario Matemático García de Galdeano* 33, 331–340.

West.D, Mangiameli P, Rampal R, and West V, 2005. Ensemble Strategies for a medical diagnosis decision support system: A breast Cancer diagnosis application. *European Journal of Operational Research*, vol.162, pp. 532-551.

Wolberg W.H, Street W.N., and Mangasarian O.L,1995. Image analysis and machine learning applied to breast cancer diagnosis and prognosis, *Analytical and Quantitative Cytology and Histology*, Vol. 17, No. 2, pp. 77-87.

Wolberg W.H., Street W.N, Heisey D.M, and Mangasarian O.L., 1995. Computerized breast cancer diagnosis and prognosis from fine needle aspirates, *Archives of Surgery*, Vol. 130, pp. 511-516.

http://www.cs.wisc.edu /~olvi/uwmp/cancer.html

http://www.ics.uci.edu/~mlearn/MLRepository.html ; also http://www.cs.wisc.edu/~olvi/uwmp/mpml.html and http://www.imagins.com/breasthealth_cancer.asp.

http://www.nd.com

http:// www.who.int

****