ISSN: 0975-766X
CODEN: IJPTFI
Research Article

Available Online through
www.ijptonline.com

CANONICAL DISCRIMINANT ANALYSIS OF STATISTICAL MODEL AND LEARNING VECTOR QUANTIZATION TECHNIQUE OF NEURAL NETWORK: A COMPARATIVE STUDY IN DIAGNOSING BREAST CANCER

[1]Dr. W. Abdul Hameed, *[2]Dr.K.Karthikeyan
[1,2]School of Advanced Sciences, VIT University, Vellore-632 014, Tamil Nadu, India
Email:k.karthikeyan@vit.ac.in

Received on 04-08-2016 *Accepted on 25-08-2016*

Abstract

Brause (2001) in his interesting study brought to limelight that human beings, however experienced and however enlightened would go wrong in diagnosing disease. The success percentage of his study are as follows:

- Best human diagnosis (most experienced Physicians): 79.7%.
- Computer with expert data base: 82.2%.
- Computer with 600 patient data: 91.1%.

This compels to establish the truth that humans cannot ad hoc analyze error-free complex data. Researchers have found that neural network capabilities can help them to improvise this domain. The implementation of human intelligence in scientific equipment has had been the subject of scientific research for a long time and of the medical research in the last decade. This paper carried out to generate and evaluate both statistical and neural network models to predict malignancy of breast tumor, using Wisconsin Diagnosis Breast Cancer Database (WDBC). The objectives in this article are: (i) Compare the diagnostic performance of statistical and neural network models in distinction between malignance and benign patterns, (ii) Reduce the number of benign cases sent for biopsy using the best model as a supportive tool, and (iii) Validate the capability of the best model to recognize new cases.

Keywords: Breast cancer, neural network, canonical discriminant analysis, learning vector quantization

1. Introduction

In the last decade, several approaches to classification had been utilized in health care applications. In order to diagnose whether the lump is benign or malignant, the Physician may use mammography, FNA (Fine Needle Aspirate) directly from the breast lump with visual interpretation or surgical biopsy. The reported ability for accurate diagnosis of cancer when the disease is prevalent is between 68% - 79%, in case of mammography (Fletcher SW *et al.*, 1993); 65% - 98% in case FNA technique is adopted (Giard RWM *et al.*, 1992), and close to 100% if a surgical biopsy is undertaken. The goal of the diagnostic aspect of this paper is to develop a relatively objective system to diagnose FNA with an accuracy that is best achieved visually.

In the past several decades, a wide variety of approaches had been proposed to achieve the perfect classification of cancerous and non cancerous cells by driving computers into the field. But the emerging technology of neural network has been largely exploited to implement a system towards classification and clustering of cells (Vombeg TW *et al.*, 2003). Artificial neural network, as a well-established computer aided diagnostic (CAD) system, is a computer algorithm capable of learning important relationships to know and to evaluate new cases. This Paper is carried out to predict malignancy tumor, using two models namely, statistical model and neural network model.

2. Material

This paper makes use of the Wisconsin Diagnosis Breast Cancer Database (WDBC) made publicly available at *http://ftp.ics.uci.edu /pub /machine-learning- database/breastcancer-wisconsin/*. This data set is the result of efforts made at the University of Wisconsin Hospital for the diagnosis of breast tumor, solely based on the Fine Needle Aspirate (FNA) test. This test involves fluid extraction from a breast mass using a small gauge needle and then a visual inspection of the fluid under a microscope. The WDBC dataset consists of 699 samples. Each sample consists of visually assessed nuclear features of FNA taken from patient's breast. Each sample has eleven attributes and each attribute has been assigned a 9-dimensional vector and is in the interval, 1 to 10 with value '1' corresponding to a normal state and '10' to the most abnormal state. Attribute '1' is sample number and attribute '11' designates whether the sample is benign or malignant. The attributes 2 to 10 are: clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, blend chromatin, normal nucleoli and mitosis. There had 16 samples that contained a single missing (i.e., unavailable) attribute value and had been removed from the database, setting apart the remaining 683 samples. Each of these 683 samples has one of two possible classes, namely benign or malignant. Out of 683 samples, 444 have been benign and the remaining 239 have been malignant, as given by WDBC dataset.

3. Neural Network Model

The powerful penetration of neural networks is due to their strong learning and generalization capability. After a neural network learns the unknown relation from given examples, it can then predict by generalization, outputs for new samples that are not included in the learning sample-set. Research on neural network dates back to the 1940s when McCulloch and Pitts found that the neuron can be modelled as a simple threshold device to perform logic function (McCulloch WS and Pitts W, 1943). In recent years, several scientific studies regard neural network and statistical techniques as meaningful alternatives to produce a useful synthesis of these two fields. There are similarities between artificial neural networks (ANNs) and statistical methods which entertain no questioning or doubting. Indeed, neural networks have been categorized as a form of nonlinear regression (Ripley BD, 1994). The multiple linear regressions which are a standard statistical tool can be expressed in terms of a simple ANN node. In plenty of cases, neural techniques and statistical techniques are seen as alternatives, or in fact, neural networks are identified as a subset of statistics. This approach has led, on one hand, to a fruitful analysis of existing neural networks and on the other hand, it has paved new avenues for current statistical methods and in few cases produced a useful synthesis of the two fields. It must be emphasized, however, that in other than some well-defined and focused application fields such as multivariate classification the two fields are different (Michie D *et al.*, 1994). This paper has been worked adopting vector quantization as a technique for breast cancer diagnosis. The basic idea is to represent the input vectors with a smaller set of prototypes that provide a congenial approximate to the input space **X**, where the vector $\mathbf{x_{i,}}$ for i =1, 2,…, N, constitutes the input space. In absence of, a priori knowledge of a probability model for the input space, it is assumed that a long training sequence of data is available.

Learning vector quantization was developed by Kohonen (1986) and in the year 1990 (Kohonen T), who summarizes three versions of the algorithm. This is a supervised learning technique that can classify the input vectors based on vector quantization. The version of LVQ incorporated in this Paper is LVQ1 - Kohonen’s first version of Learning Vector Quantization (Kohonen T, 1986). The set of input vectors have been denoted as $\{x_i\}$, for i =1, 2,…,N, and the network synaptic input vectors has been denoted as $\{w_j\}$, for j = 1, 2, …,m. C_{w_j} has been the class (or category) that has been associated with the (weight) Voronoi vector $w_{j,}$ and C_{xi} as the class label of the input vector x_i to the network. The weight vector w_j has been adjusted in the following manner:

If the class associated with the weight vector and the class label of the input are the same, that is, $Cw_j = Cx_i$, then

$$w_j(k + 1) = w_j(k) + \mu(k) [x_i - w_j(k)] \qquad (1)$$

where $0< \mu(k)<1$ (the learning rate parameter).

But if $Cw_j \neq Cx_i$, then

$$w_j(k + 1) = w_j(k) - \mu(k) [x_i - w_j(k)] \qquad (2)$$

and the other weight vectors are not adopted. Therefore, the update rule for modifying a weight vector in (1) is absolutely standard, if the class is correct. In other words, according to the learning rule in (1), the weight vector w_j is moved in the direction of the input $\mathbf{x_i}$, if the class labels of the input vector and of the weight vector agree. However, if the class is not correct, the weight vector is moved in the opposite direction away from the input vector, according to (2).

Algorithm:

The basic LVQ1 algorithm can be summarized as follows:

Step 1: Initialize all weight vectors $w_j(0)$, initialize the learning- rate parameter $\mu(0)$ and set $k = 0$.

Step 2: Check the stopping condition. If false, continue; if true, quit.

Step 3: For each training vector $\mathbf{x_i}$, perform step 4 and step 5.

Step 4: Determine the weight vector index (j=q) such that min $\|x_i - w_j(k)\|^2$ and the weight vector $w_q(k)$ that minimizes the square of the norm

Step 5: Update the appropriate weight vector w_q (k) as follows:

If $Cw_q = Cx_i$, then $w_q (k+1) = w_q (k) + \mu (k) [x_i - w_q (k)]$

If $Cw_q \neq Cx_i$, then $w_q (k+1) = w_q (k) - \mu (k) [x_i - w_q (k)]$

Step 6: Set k ← k+1, reduce the learning rate parameter, and then go to step 2. The learning rate parameter **μ** can be reduced in accordance with k (discrete time index) using $\mu(k) = \mu(k+1)/ k+1$ for $k > 0$.

4. Statistical Model

Canonical Discriminant Analysis (CDA) is a statistical model and a multivariate method used for demonstrating the significance and nature of the differences between two or more pre-defined groups of objects where data are available for several variables measured on each object (Johnson RA and Wichern DW, 1982). The main objectives of this analysis are: (i) to find a set of discriminant functions with power in decreasing order of discrimination between groups identified as a priori, (ii) to test whether the means of these groups along that axis are significantly different, and (iii) to attempt to assign individual objects of unknown origin to the given groups. Given two or more groups of observations with measurements on several interval variables, canonical discriminant analysis derives a linear combination of the variables that has the highest possible multiple correlation with the groups. This maximal multiple correlation is called the first canonical correlation. The coefficients of the linear combination are the canonical coefficients or canonical weights. The variable defined by the linear combination is the first canonical variable or canonical component. The second canonical correlation is obtained by finding the linear combination uncorrelated with the first canonical variable that have the highest possible multiple correlation with the groups. The process of extracting canonical variables can be repeated until the number of canonical variables equals the number of original variables or the number of classes minus one, whichever is smaller. The first canonical correlation is at least as large as the multiple correlation between the groups and any of the original variables. If the original variables have high within-group correlations, the first canonical correlation can be large even if all the multiple correlations are small. In other words, the first canonical variable can show substantial differences among the classes, even if none of the original variables does. For each canonical correlation, canonical discriminant analysis tests the hypothesis that it and all smaller canonical correlations are zero in the population. An F approximation is used that gives better small-sample results than the usual χ^2 approximation. The variables should have an approximate multivariate normal distribution within each class, with a common covariance matrix in order for the probability levels to be valid. The new variables with canonical variable scores in canonical discriminant analysis have either pooled within-class variances equal to one or total-sample variances equal to one.

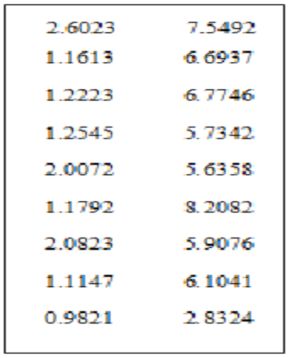
5. Methods

This paper has compared two models namely, Neural Network model using the Kohonen's first version of Learning Vector Quantization technique and statistical model using the canonical discriminant analysis for breast cancer diagnosis. To determine the performance of these models in practical usage, the database has been divided randomly into two separate sets, one for training and another for validation: (a) the training samples comprising **500** patients records (**303** benign, **197** malignant) and (b) the validation samples comprising **183** patients records (**141** benign, **42** malignant). Using the patients' records in training sample the models have been trained by adjusting the weight values for interconnection limits for the neural network model. Then, the patients' record in validation samples (n = 183) have been utilized to evaluate the generalizing ability of the above said models, separately. The best of these models has been compared in terms of accuracy, sensitivity, specificity, false positive and false negative.

6. Experiment and Results

Out of **683** samples of Wisconsin Diagnosis Breast Cancer data set, **500** samples have been used to train the neural network using the Kohonen's first version of Learning Vector Quantization (LVQ1) technique and the remaining **183** samples have been used to test the sample data. The Matlab software has been used with the number of training epochs set to **35,000** to find the codebook vectors. The final weight in table-1 presents the codebook vectors, after 35,000 training epochs. Table-2, presents the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) results. Thus, having a priori known set of **444** benign and **239** malignant instances, the neural network has successfully identified **435** (**96.67%**) instances as negative and **224** (**96.14%**) instances as positive.

Table-1: Code book vectors after 35,000 training epochs.



2.6023	7.5492
1.1613	6.6937
1.2223	6.7746
1.2545	5.7342
2.0072	5.6358
1.1792	8.2082
2.0823	5.9076
1.1147	6.1041
0.9821	2.8324

According to these observations, Table-3 presents the sensitivity, specificity and efficiency of the neural network model using LVQ1 technique, as well as the predicted values of a positive/negative test results.

As a second model, the Researcher has applied Statistical Model namely the Canonical Discriminant Analysis method for the same 683 samples of WDBC data set. Table 2 presents the results that have been obtained by using this model. Out of **444** benign and **239** malignant instances, the canonical discriminant analysis has successfully identified **436 (96.25%)** instances as negative and **222 (96.52%)** instances as positive. In the same table are also presented the true positive (TP), false positive (FP), true negative (TN) and false negative (FN) results. In Table-3, the sensitivity, specificity and efficiency of the canonical discriminant analysis, as well as the predicted values of positive/negative test results are recorded.

Table-2: Comparative performance of the models.

Data	Group	Actual Instances	Predicated instances			
			Neural Network		Canonical discriminant	
			Positive	Negative	Positive	Negative
Training Data (500)	Benign	303	09 (FP)	294 (TN)	08 (FP)	295 (TN)
	Malignant	197	182 (TP)	15 (FN)	181 (TP)	16 (FN)
Testing Data (183)	Benign	141	0 (FP)	141 (TN)	0 (FP)	141 (TN)
	Malignant	42	42 (TP)	0 (FN)	42 (TP)	0 (FN)
Total Data (683)	Benign	444	09 (FP)	435 (TN)	8 (FP)	436 (TN)
	Malignant	239	224 (TP)	15 (FN)	222 (TP)	17 (FN)

Table-3: Comparative results of the models.

Data	Measurements	Neural Network	Canonical discriminant
Training Data (500)	Sensitivity Specificity Efficiency Predictive value (Positive) Predictive value (Negative)	92.39 97.03 95.20 95.29 95.15	91.88 97.36 95.20 95.77 94.86
Testing Data (183)	Sensitivity Specificity Efficiency Predictive value (Positive) Predictive value (Negative)	100 100 100 100 100	100 100 100 100 100
Total Data (683)	Sensitivity Specificity Efficiency Predictive value (Positive) Predictive value (Negative)	93.72 97.97 **96.49** 96.14 **96.67**	92.89 98.20 **96.34** 96.52 **96.25**

7. Conclusions

This paper has compared two models namely neural Network model using the Kohonen's first version of Learning Vector Quantization technique and statistical model using the canonical discriminant analysis for the diagnosis of breast cancer. The ability of these models to differentiate malignant from benign tumor the Researcher has compared a group of **683** patients. The main aim of this paper is to investigate which model obtains more reasonable specificity while keeping high sensitivity. The benefit is that the number of breast cancer patients for biopsy can be restricted. The seriousness of the ailment can easily be assessed. The output of the artificial neural network model has yielded a perfect sensitivity of **93.72%,** specificity of **97.97%** and high efficiency of **96.49%** for the total data set. The output of the canonical discriminant analysis (CDA) has yielded a sensitivity of **92.89%**, maximum specificity of **98.20%** and efficiency of **96.34%** for the total data set, exhibiting the fact that canonical discriminant analysis also results in similar proportions as ANN to differentiate malignant and benign tumors. We developed a diagnostic system that performs at or above an accuracy level in any procedure short of surgery. The results have also suggested that neural network model and statistical model are a potentially useful multivariate method for optimizing the diagnostic validity of laboratory data. The physicians can combine this unique opportunity extended by statistical and neural network models with their expertise to detect the early stages of the disease.

References

1. Brause RW (2001), "Medical Analysis and Diagnosis by Neural Networks", Computer Science Department, Fran furt, A. M., Germany
2. Fletcher SW, Black W, Harrier R, Rimer BK and Shapiro S (1993), "Report of the International workshop on screening for breast cancer", Journal of the National Cancer Institute, 85:1644-1656
3. Giard RWM and Hermann J (1992), "The value of aspiration cytologic examination of the breast: A statistical review of the medical literature", Cancer, 69:2104-2110
4. Johnson RA and Wichern DW, "Applied Multivariate Statistical Analysis", Prentice-Hall, Inc., Englewood Cliffs, New Jersey, 1982.
5. Kohonen T (1986), "Learning Vector Quantization for Pattern Recognition", Technical Report TKK-F-A601, Helsinki University of Technology, Finland.
6. Kohonen T (1990), "Improved Version of Learning Vector Quantization", Proceedings of the International Joint Conference on Neural Networks, San Diego, CA, vol. 1, pp.545-550
7. Michie D, Spiegel halter DJ, Taylor CC, (1994) Eds., "Machine Learning, Neural, and Statistical Classification", London: Ellis Harwood Ltd.
8. McCuulloch WS and Pitts W (1943), "A logical calculus of the ideas immanent in nervous activity", Bull. Of mathematical Biophysics, 5:115-133
9. Vombeg TW, Buscema M, Kanczer HU, Teifke A, Intraliga M, Terzi S, Heussel P, Achenbach T, Ricker O, Mayer D and Thelen M (2003), "Improved artificial neural network in prediction of malignancy of lesions in contrast – enhanced MR-mammography", Med. Phys., 30:2350-2359.
10. Ripley B (1994), "Neural networks and related methods for classification", J. Roy. Stat. Soc., Series B, 56(3):409-456

Corresponding Author:
Dr.K.Karthikeyan*,
Email: k.karthikeyan@vit.ac.in