*International Journal of Engineering & Technology*, 7 (4.10) (2018) 922-924

International Journal of Engineering & Technology

*Website: www.sciencepubco.com/index.php/IJET*

***Research paper***



# Breast Cancer Prognosis Using Learning Vector Quantization Neural Network Technique

**W. Abdul Hameed[1], Raja Das[1*], Jitendra Jaiswal[1]**

*[1]School of Advanced Sciences, VIT University, Vellore-632 014, Tamil Nadu, India*
**Corresponding author E-mail: rdasresearch@gmail.com*

**Abstract**

A suitable treatment coming after surgery is very much motivated by prognosis - the speculated outcome of the disease. Now-a-days improving prognostic prediction is a challenging task to the doctors. This paper presents prognosis for the breast cancer issues by applying Neural Network Architecture with the dataset for Wisconsin Prognostic Breast Cancer. The accuracy is evaluated by adopting algorithm for Kohonen's first issue of Learning Vector Quantization to predict the recurrence of the disease within 2 years or beyond and also within 5 years or beyond.

***Keywords***: *Neural networks; Learning Vector Quantization; Breast cancer Prognosis; recurrence / non-recurrence; Classification.*

## 1. Introduction

Breast Cancer is one of the most serious issues from common cancer for Indian women. It is quite difficult for making a prediction as to whether or not a malign tumor will be repeated is by nature. Many researches have been continued with a diversification of machine-learning techniques to accomplish this task, including decision trees [1], separation based on medians [2], and artificial neural networks [3] (Wolberg, 1994; Street, 2000; Repley, 1998).
Other groups also extended their research for prognosis using different method, such as feed forward neural network using back-propagation algorithm [4], entropy maximization networks [5, 6], and fuzzy-based techniques [7]. In conventional medical researches, the dependable prognosis measure is to probe the dimensions to which cancer is prevailed in the lymph nodes [8]. It necessitates the surgical diminishing of the nodes, enabling the patients' immune to infections. Contemporary research is to obtain directly from the tumor mass to predict breast cancer recurrence [9].

## 2. Material and Method

### 2.1. Material

The breast cancer prognosis phenomena with the Wisconsin Prognosis Breast Cancer (WPBC) dataset has been addressed in this paper, which is publicly available via an anonymous ftp [10]. The WPBC dataset is the outcome of attempts, carried out at the University of Wisconsin Hospital for the prognosis of breast tumour solely based on the fine needle aspirate (FNA) test. The WPBC data consists of 198 instances (47 recurrences and 151 non-recurrences), where each entity represents patronizing data for single breast cancer case. It was consecutive record for patients at the Wisconsin University Hospital for the tenure of 1984 to 1995 and including only those cases which exhibit invasive metastases at the time of diagnosis.

WPBC Cell Nuclei characteristics (or columns) are compactness, radius, perimeter, texture, area, concavity, smoothness, concave points, symmetry, and fractal dimension. This data had been applied for many medical literatures [11].

### 2.2 Artificial Neural Network

In the year 1951, the definition of the first artificial neuron [12] was defined by McCullogh and Pitts. Based on the works of McLelland, Rummelhart, Hopfield and Kohohnen [13,14,15] during the period 1982 to 1987, Mathematical models were applied for practical applications.
Artificial Neural network are based on calculated paradigm with mathematical models, those unlike conventional computing have an operation and structure of the mammal brain.
An important motivation for neural network is to shape the computational properties and functional characteristic of the brain when it executes cognitive processes such as concept, sensorial perceptions, categorisation, learning and concept association. In recent years, they have been used to analyse data from a variety of human clinical studies. ANN operates in two different modes: training (or learning) and testing. Training is simply an adaptive process which follows the weights assigned to interconnected neurons in order to access the most fitting outcome for all the observed stimuli however testing assures the performance of the algorithm.

### 2.3 Training Method

Vector quantization is a neural network method used for clustering and classification. The core concept is to replace the already assigned input weight vectors with a reduced set of prototypes that provides a closest approximation to the input space *Y*, where the vector $y_i$ for $i = 1, 2, \cdots, n$, form the input space. It has been considered that no prior knowledge of a probability model for the input space is given; nevertheless, it is assumed that a long train-

**Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.**

ing sequence of data is available. The objective of this research is to develop a "code book" for quantization vectors and then is to encode any input vector.

To develop a codebook, a large set of training vectors is used to form groups according to a predetermined number of clusters and each cluster '*j*' is represented by its particular centroid $y_i^*$ (i.e., the reproduction or reconstruction vector). The centroid clustering is based on a given distortion measure. The distortion method which is used here is based on the L2 (Euclidean) norm of a vector, that is

$$d\left(y, y^*\right) = \left\|y - y^*\right\|^2 = \left(y - y^*\right)^T \left(y - y^*\right) \quad (1)$$

Referred to as the squared-error distortion. Once the codebook is created, it is stored in the "transmitter" and the "receiver". The quantization of an input vector $y_i$ is then performed as follows:

1. The input is given to the vector quantizer and is compared to the codebook $y_j^*$ for $j = 1, 2, \cdots, m$, and the codebook vector that gives the minimum distortion, that is, the minimum distance according to (1) is selected.
2. The selected vector $y_q^*$ is then represented by $y_i$ and the index *q* (associated with the appropriate class to which the input vector belongs) is "transmitted" to the receiver where the appropriate reproduction (reconstruction) vector $y_q^*$ is selected as the representation of the input vector $y_i$. When the Euclidean distance (distortion) measure is used to decide to which region the input vector $y_i$ belongs, and the quantizer is called as a Voronoi quantizer.

The Voronoi quantizer performs a partitioning of the input space into various Voronoi cells, where each cell is represented by one of the reproduction vectors $y_j^*$. Then the $q^{\text{th}}$ Voronoi cell then contains those points of the input space Y that are closer to the reconstruction vector $y_q^*$, in the Euclidean sense, than to any other reproduction vector $y_j^* \left(i \neq q\right)$.

In this paper we have applied Kohonen's first phase of learning vector quantization (LVQ1). It sets in motion with an input vector chosen randomly from the training dataset. The output neuron OR the class in LVQ1 is declared as "winner" according to

$$\min d\left(y_i, w_j\right) = \min \left\|y_i - w_j\right\|^2 \quad (2)$$

where the minimum is taken over all . In (2), the synaptic weight vector $w_j$ has replaced the reproduction vector $y^*$ shown in (1).

### 2.4 Algorithm

$\left\{y_i\right\}, i = 1, 2, \cdots, n$ be the input.

$\left\{w_j\right\}, j = 1, 2, \cdots, m$ be the synaptic weight

$Cw_j$ be the class associated with $w_j$

$Cy_i$ be the class associated with $y_i$

The vector $w_j$ is updated in the following steps:

If $Cw_j = Cy_i$, then

$$w_j\left(k+1\right) = w_j\left(k\right) + \mu\left(k\right)\left[y_i - w_j\left(k\right)\right] \quad (3)$$

Where $\mu$ is the learning rate parameter.

But if $Cw_j \neq Cy_i$, then

$$w_j\left(k+1\right) = w_j\left(k\right) - \mu\left(k\right)\left[y_i - w_j\left(k\right)\right] \quad (4)$$

and the other weight vectors are not adapted.

This research considers that the first $m$ (total # of classes) vectors from learning data set are used to initialize the weight vectors, that is, $w_j\left(0\right)$ for $j = 1, 2, \cdots, m$. The terminating condition control can be following the total number of training iterations.

## 3. Experiment and Result

The dataset from Wisconsin Prognosis Breast Cancer (WPBC) has been considered for this experiment. It has 198 instances in which 47 have been recurrences and 151 have been non-recurrences, where each has represented sequenced data for one breast cancer case. Out of 198 instances, 27 have been recurrence instances which occur within 2 years and 109 have been non-recurrence instances which occur beyond 2 years. This subset of data having 27 recurrence and 109 non-recurrence instances have been taken together to our experiment to develop a codebook vector and it has been used for the new cancer patients to predict whether the disease recur within 2 years or beyond. The topology of the neural network for 2 years prognosis is 32-136-2. There are 32 nodes in the input layer, where each node (attributes are considered as nodes) communicates to the cell nuclei features, which is given in Table 1. The proceeding or second layer comprises 136 nodes, which communicates to the total amount of instances, namely 27 recurrence and 109 non-recurrence instances, for the training epoch. Eventually, the summation/division layer comprises 2 nodes that represent the classification namely, the recurrence of the disease before 2 years or beyond. The accuracy of the prediction using the Kohonen's first phase of Learning Vector Quantization (LVQ1) technique for the WPBC dataset is depicted in Table I.

**Table I:** Performance of LVQ1 to Predict the Disease to recur in 2 years or beyond

| Recurrence/ Non-recurrence | Actual instances | Predicted instance | % of correctly predicted instances |
|---|---|---|---|
| Recurrence Before 2 years | 27 | 21 | 78 % |
| Recurrence Beyond 2 years | 109 | 67 | 62 % |
| Total | 136 | 88 | 65 % |

Similarly for 5 years prognosis, the topology is 32-100-2. This subset of data having 40 recurrence instances, which occur within 5 years and 60 non-recurrence instances that occur beyond 5 years have been taken together for our experimentation to develop a codebook and it has been used for the new cancer patients to predict whether the disease recur within 5 years or beyond. The accuracy of the prediction for this dataset is presented in Table II.

**Table II:** Performance of LVQ1 to Predict the Disease to recur in 5 years or Beyond

| **Recurrence/ Non-** | **Actual instances** | **Predicted instance** | **% of correctly predicted instances** |
|---|---|---|---|

| recurrence | | | |
|---|---|---|---|
| Recurrence Before 5 years | 40 | 31 | 78 % |
| Recurrence Beyond 5 years | 60 | 41 | 68 % |
| Total | 100 | 72 | 72 % |

## 4. Conclusion

This study has shown that LVQ1 technique has correctly predicted the prognosis of the breast cancer to an accuracy of 65 % for 2 years and 72 % for 5 years. In the cases when some patients will impatiently change hospitals, or doctors, or die of diseases unrelated to the cancer, the right endpoints of the recurrence time intervals are censored. Therefore, the training data for the learning phase may not be always well-defined. Moreover, the neural network works perfectly only if the training dataset is large enough to learn the problem. In this problem, since the training dataset is very small, the accuracy rate is not high.

The performance in prognostic diagnosis made possible by neural networks may be clinically contributing for clinical trials, therapy, patient information, and quality assurance. In the action of making important decisions regarding therapy, the prognostic diagnosis may permit the efficient segregation of patients with a delicate prognosis (who require therapy) from patients with a proficient prognosis (who need little or no therapy), and it may also predict who will accurately respond to a particular therapy. This would empower for the initiation of more homogenous patient population for clinical experiments, resulting in fewer clinical trial patient population. Moreover, it canalizes less expensive trials, and the potential to probe treatment consequences that may be unnoticeable in more rigorous study population. Pertaining to patient information, our outcomes may provide more clear eyesight to the patients for their disease. Ultimately, for estimation and quality undertaking, they may consider a better seriousness of illness reconciliation.

## References


[1] W. H. Wolberg, W. N. Street, and O. L. Mangasarian (1994), Machine Learning Techniques to diagnose Breast Cancer from Image-processed Nuclear Features of Fine Needle Aspirates, *Cancer Letters*, vol. 77, pp. 163-171.

[2] W. N. Street (2000), Xcyt: A system for remote cytological diagnosis and prognosis of breast cancer, soft-computing Techniques in Breast Cancer Prognosis and Diagnosis, *World Scientific Publications*.

[3] R. M. Ripley (1998), Neural network for Breast cancer Prognosis, *Ph.D., Thesis, Department of Engineering Science, University of Oxford*.

[4] H. B. Burke, and P. H. Goodman (1997), Artificial neural network improve the accuracy of cancer survival prediction, *Cancer Journal*, vol. 99, pp. 857-862.

[5] P. L. Choong, and C. J. S. deSilva (1996), Entropy maximization networks, An application to breast cancer prognosis, *IEEE Transaction Neural Network*, vol. 7, pp. 568-577.

[6] P. L. Choong, and C. J. S. deSilva (1998), Maximum entropy estimation vs. multivariate logistic regression: Which should be used for the analysis of small binary outcome data sets ?, *Proc. 20th Ann. Int. Conf. IEEE Eng. Med. Biol. Soc.*, vol. 3, pp. 1602-1605.

[7] H. Sekar, M. Odetayo, D. Petrovic, R. N. Naguib, C. Bartolic, L. Alasio, M. S. Lakshmi and G. V. Sherbet (2000), A fuzzy measurement based assessment of breast cancer prognostic markers, *Proc. Of 2000 IEEE EMBS Int. conf. on Information Tech. Applications in Biomedicine*, pp174-178.

[8] O. L. Mangasarian, W. N. Street, and W. H. Wolberg (1995), Breast cancer diagnosis and prognosis via linear programming, *Operations Research*, vol. 43, no. 4, pp. 570-577.

[9] W. N. Street (1994), Cancer diagnosis and prognosis via linear programming based machine learning, *Ph.D. dissertation, University of Wisconsin*.

[10] http://ftp.ics.uci.edu/pub/machine-learning-databases/breast cancer-Wisconsin/Wisconsin Diagnosis Breast Cancer (WDBC) Dataset and Wisconsin Prognostic Breast Cancer (WPBC) Dataset

[11] W. H. Wolberg, SW. N. Street, and O. L. Mangasarian (1995), Image analysis and machine learning applied to breast cancer diagnosis and prognosis, *Anal Quant cytol* , vol. 17, pp. 77-87.

[12] W. S. McCullogh (1943), W. Pitts, Bull Maths Biophysics, vol. 5, pp. 115-118.

[13] D. E. Rummelhart, J. L. McLelland (1987), Parallel distributed processing: Explorations in the microstructure of Cognition, *Foundation, MIT Press*, vol. I.

[14] J. J. Hopfield (1982), Neural networks and physical systems with emergent collective computational abilities, *Proc. Nat. Acad. Sci*. pp. 2554-2558.

[15] T. Kohonen (1982), Self-organized formation of topologically correct feature maps, *Biological cybernetics*, vol. 43, pp. 59-69.