Breast Cancer Diagnostic System using Hierarchical Learning Vector Quantization

R.R.Janghel
Soft Computing & Expert System Laboratory
ABV-IIITM, Gwalior, India

Ritu Tiwari
Soft Computing & Expert System Laboratory
ABV-IIITM, Gwalior, India

Anupam Shukla
Soft Computing & Expert System Laboratory
ABV-IIITM, Gwalior, India

ABSTRACT
Breast cancer has become a common mortality factor in the world. Lesser availability of diagnostic facilities along with large time requirements in manual diagnosis emphasize on automatic diagnosis for early diagnosis of the disease. In this paper a computerized breast cancer diagnosis prototype has been developed to reduce the time taken and indirectly reducing the probability of death. The paper presents Hierarchical Learning Vector Quantization (HLVQ) as a classifier for the diagnosis. Hierarchical LVQ networks consist of multiple LVQ networks assembled in different level or cascade architecture. In this research two stage of LVQ network is used on WDBC datasets. The first level of LVQ reduces the feature space which is further worked over by the second stage for computing the output. The experiments confirm an effective detection of the disease by use of multiple networks. A comparative study of work carried in the field of breast cancer diagnosis using different ANN algorithm is also done.

Keywords
Breast cancer, neural networks, diagnosis, Learning Vector Quantization Hierarchical LVQ

1. INTRODUCTION
Carcinoma of the breast is a second reason after lung cancer as a tumor-related cause of death in women (10.4% of all cancer incidence, both sexes counted) [1] and the fifth most common cause of cancer death [2]. In 2004, breast cancer caused 519,000 deaths worldwide (7% of cancer deaths; almost 1% of all deaths) [2]. Currently, more than 190,000 new cases and 40,000 deaths occur in the US alone [3]. Carcinoma generally begins as a focal curable disease, but usually is not identifiable at this stage by the physician during examination. It is estimated that the mortality from breast carcinoma could be decreased by up to 25% provided that all in the appropriate age groups were regularly screened [4].

Breast cancer is the most common form of cancer in women in the US and Europe. Diagnosis is generally provided by a mammogram, with additional support provided by cytological examination, and the experience and opinion of the attending surgeon. In the case of a positive diagnosis, in many cases a radical mastectomy is to be performed. It is critical that the accuracy of the diagnosis is 100% - or as close as humanly possible because the treatment for the disease can be permanent and drastic (i.e. radical mastectomy). Therefore, most medical institutions will insist that their personnel err on the side of caution - minimizing the risk by maximizing the specificity of the diagnosis. Generally these results in a reduction in sensitivity as many patients with symptoms falling on the tails of the distribution will not be properly diagnosed. Sensitivity reflects the level of false negatives, which must be below acceptable levels - on the order of less than 1% is desirable. How this result is to be achieved is still uncertain [5]. In this, we use a well structured and complete breast cancer dataset that has been used as a bench mark test for various machine learning techniques. The database is worked over by a new proposed method of classification that makes use of a couple of LVQ networks in a layered mode.

The breast cancer datasets used in this paper were generated using the Xcyt image analysis program [6, 7]. First, a sample of fluid is taken from the patient's breast by a fine needle aspirate. An image of the fluid is transferred to a workstation by a video camera mounted on a microscope. Xcyt is a system that provides expert diagnosis and prognosis of breast cancer based on fine needle aspirates. The system combines techniques of digital image analysis, inductive machine learning, mathematical programming, and statistics, including novel prediction methods developed specifically to make best use of the cytological data available. The result is a program that diagnoses breast masses with an accuracy of over 97%, and predicts recurrence of malignant samples without requiring lymph node extraction [8-18]. The accuracy may be further improved by better classifiers which is the focus of the present paper.

2. METHODS AND PROCEDURES
As the sophistication of Artificial Neural Networks (ANNs) and other learning systems are enhanced, their applications become increasingly diverse. The ability to provide real-time results coupled with almost limitless configurability makes ANNs a good candidate for the statistical detection of breast cancer.

In this paper we develop a hierarchical LVQ model and use it for diagnosis of breast cancer. The block diagram of the methodology used is shown in the figure 1. The data for the problem is taken from the Winconsin Breast Cancer Dataset. The dataset has some missing values for which the corresponding data entry is removed in the pre-processing stage. The resultant complete dataset is divided into training and testing sub-sets used for training and testing respectively.

For discussion we first discuss the basic LVQ and then the proposed hierarchical LVQ.

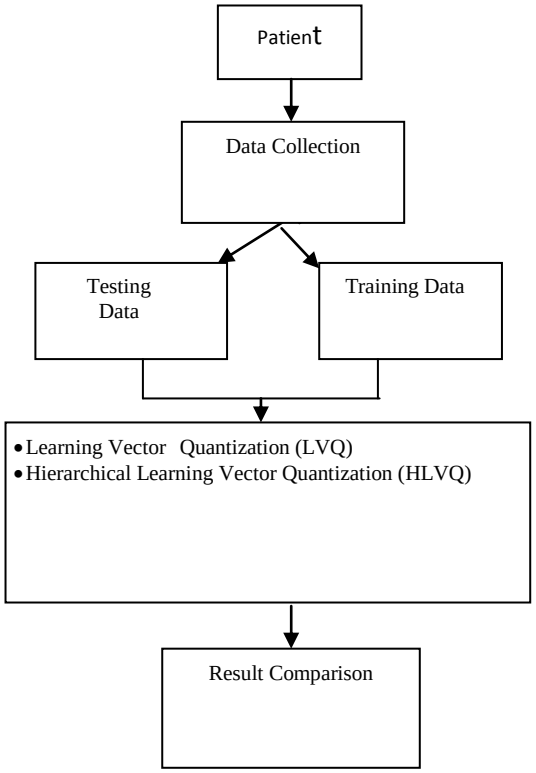

Figure 1 block diagram of overall methodology

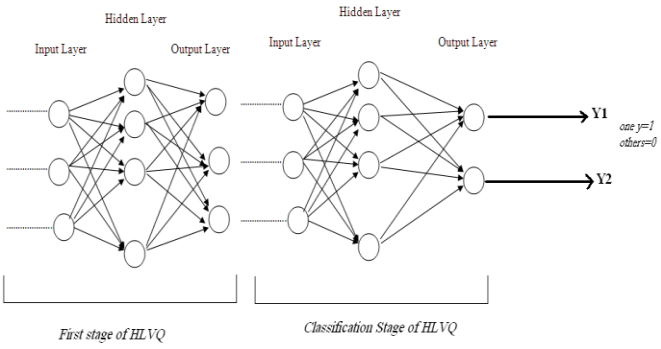

Figure 2: Architecture of HLVQ

2.1 Learning Vector Quantization (LVQ):

LVQ is an adaptive data classification method based on training data with desired class information. Although a supervised training method, LVQ employs unsupervised data clustering techniques (e.g., competitive learning) to preprocess the data set and obtain cluster centers.

The general concept of the LVQs comes from the Hebbian Learning. The Hebbian Learning is an understanding of the human brain and the learning associated with it. The Hebbian Learning is derived from the Hebb's postulate which states that "When an axon of cell A is near enough to excite a cell B and repeatedly or persistently takes part in firing it, some growth processes or metabolic changes takes place in one or both cells such that A's efficiency, as one of the cells firing B, is increased."

The approach is also termed as the winner-takes-all approach. This is because we first study the activities of all the neurons. Then we decide the winning neuron. The activity of this winning neuron is adjusted according to the answer. This may be visualized to be a very apt learning strategy for any classificatory problem where the winning neuron is the point in the input space and the final output is nothing but the class to which it belongs to. The continuous adjustment of the neuron activity and their effects on the neighboring neuronal activity in multiple epochs is done. This adjusts the various weights or parameters of the network in order to better adapt the network to the problem. This makes these networks perform very well for the classificatory problems.

The basic LVQ algorithm is the following:

1) Initialize the codebook vectors Wi and the learning rate α
2) Randomly select an input vector X
3) Find the winner unit closest to the input vector (i.e. the codebook vector Wc with the smallest Euclidean distance with regard to the input vector X):

$$\|X - W_c\| = \min_k \|X - W_k\| \quad \text{i.e;}$$

$$c = \arg\min_k \|X - W_k\|.$$

4) Modify the weights of the winner unit:
   - If W_c and X belong to the same class (the classification has been correct) $W_c(t+1)=W_c(t)+\alpha(t)[X(t)-W_c(t)]$.
   - If W_c and X belong to different classes (the classification has not been correct) $W_c(t+1)=W_c(t)-\alpha(t)[X(t)-W_c(t)]$.
5) Reduce the learning rate α
6) Repeat from step 2 until the neural network is stabilized or until a fixed number of iterations have been carried out

2.2 Hierarchical Learning Vector Quantization:-

This paper proposes HLVQ algorithm which reduces the limitations of the primitive LVQ algorithm. The proposed method is suitable for large scale classification problems. The general architecture of the system is given in figure 2.

The proposed system consists of two stage network. LVQ is used in the first stage to reduce the candidate categories quickly. In the second stage, another LVQ network is used for classify between benign or malignant which are the two possible output classes. As and when the need arises, nodes are multiplied hierarchically and unnecessary codebook vectors are removed by the original criteria based on the number of categories and vectors.

HLVQ tackles this problem by dividing the feature space into areas which overlap at the borders. HLVQ divides the feature space and separates the categories hierarchically using a few codebook vectors. In learning, the structure of network is adapted by the original criteria based on the number of categories and vectors. Since the feature space with a large amount of categories has complicated borders, HLVQ divides the feature space into regions with overlapping areas.

In this paper a Breast Cancer Diagnosis System using Hierarchical Learning Vector Quantization (HLVQ) is proposed. The proposed system consists of two stage network. LVQ used in the first stage prunes the candidate categories quickly. In the second stage, second LVQ is used for distributing the input into final class under HLVQ.

Algorithm for HLVQ

(Let C be First stage of HLVQ, D be Second Stage of HLVQ)

Initialize the codebook vectors Wi and the learning rate α for C

Randomly select an input vector X1

Find the winner unit closest to the input vector (i.e. the codebook vector Wc with the smallest Euclidean distance with regard to the input vector X):

i.e.

Modify the weights of the winner unit:

If Wc and X belong to the same class (the classification has been correct)

$Wc(t+1)=Wc(t)+\alpha(t)[X(t)-Wc(t)]$.

If Wc and X belong to different classes (the classification has not been correct)

$Wc(t+1)=Wc(t)-\alpha(t)[X(t)-Wc(t)]$.

Repeat from step 2 until the neural network C is stabilized or until a fixed number of iterations have been carried out.

Initialize the codebook vectors Wj for D

Randomly select an input vector X2

Find the winner unit closest to the input vector (i.e. the codebook vector Wc with the smallest Euclidean distance with regard to the input vector X):

Modify the weights of the winner unit:

If Wc and X belong to the same class (the classification has been correct)

$Wc(t+1)=Wc(t)+\alpha(t)[X(t)-Wc(t)]$.

If Wc and X belong to different classes (the classification has not been correct)

$Wc(t+1)=Wc(t)-\alpha(t)[X(t)-Wc(t)]$.

Reduce the learning rate α

Repeat from step 7 until the neural network D is stabilized or until a fixed number of iterations have been carried out.

The basic working methodology, as given in figure 1, consists of the steps of data set collection, pre-processing, and training along with testing.

The first step involves data set collection. Wisconsin Breast Cancer datasets are collected from various sources. Features are computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. They describe characteristics of the cell nuclei present in the image. In the datasets, ten features are computed for each nucleus: radius, texture (variance of grey levels inside the boundary), perimeter, area, smoothness (local variation of radial segments), compactness, fractal dimension (of the boundary), symmetry, concavity, and the number of concave points. The mean value, extreme value, and standard error of these features for all cells are also computed, resulting in a total of 30 morphometric features for each patient.

Attribute Information are Diagnosis (M = malignant, B = benign),3-32:- Ten real-valued features are computed for each cell nucleus, radius (mean of distances from center to points on the perimeter),texture (standard deviation of gray-scale values),perimeter, area, smoothness (local variation in radius lengths),compactness (perimeter2/ area - 1.0),concavity (severity of concave portions of the contour),concave points (number of concave portions of the contour),symmetry ,fractal dimension ("coastline approximation" - 1).The mean, standard error, and "worst" or largest (mean of the three largest values) of these features were computed for each image, resulting in 30 features. For instance, field 3 is Mean Radius, field 13 is Radius SE, and field 23 is Worst Radius.

All feature values are recoded with four significant digits. Dataset consists of 357 benign cases and 212 malignant cases.

Benign tumors generally have smooth, circumscribed, and well-defined contours, whereas malignant tumors commonly have rough, speculated, and ill-defined contours. Based upon this observation, several shape factors have been developed. The boundaries of typical malignant tumors are more irregular than those of typical benign ones obviously. Thus, it is feasible to distinguish malignant tumors from benign ones through contour complexity feature.

After data collection and feature extraction, preprocessing is performed on the available data. In this first, instances with missing attribute values are removed. Then the data is divided into three sample of different number of training set and testing set.

Training set contains training data and testing set consist data of to be used for the testing purposes. Each of the data sets has cases of benign and malignant. The details of the three samples are as follows.

Sample 1:

Training Set:-it consist data of first 400 patents, malignant cases – 173, benign cases – 227, Testing Set: - it consist data of another 169 patents. Malignant cases – 39, Benign cases – 130

Sample 2:

Training Set:-it consist data of first 250 patents. Malignant cases – 124, benign cases – 126, Testing Set: - it consist data of another 250 patents. Malignant cases are 71, Benign cases – 179.

Sample 3:

Training Set:-it consist data of first 169 patents. Malignant cases – 89, benign cases – 80, Testing Set: - it consist data of another 400 patents. Malignant cases – 123 benign cases – 277

After dividing the datasets into training set and testing set. These sets are used to train and test the formulated Hierarchical Learning Vector Quantization Neural Network.

First the network is trained with training set of all samples individually. Then first this trained NN is tested using training set and the accuracy is calculated.

3. EXPERIMENTAL RESULTS

Into a scenario where medical diagnosis is gaining increasing importance, there is a certain need to introduce computational intelligence techniques into the traditional bio-medical domain. This enables making accurate decision in the least possible time. The present paper talks about the use of hierarchical LVQ network for the diagnosis of breast cancer. The approach made use of two LVQ networks, one after the other, for diagnosis. On the basis of the experimental

observations of different neural network, HLVQ gives the highest accuracy of 98.14% for diagnosis of breast cancer.

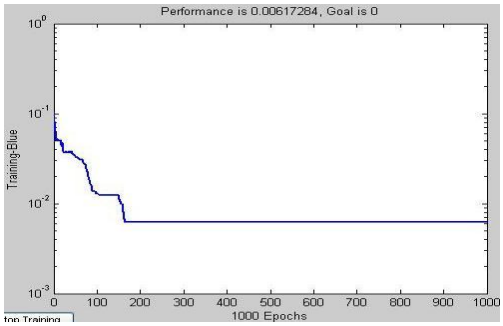

Figure 3: Result after first stage of HLVQ

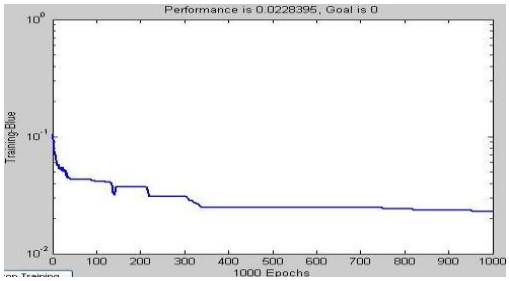

Figure 4: Result after second stage of HLVQ

Table 1: Performance comparison of Neural networks models

ANN Algorithm	Accuracy %
HLVQ	**98.14**
BPA	97.50
RBF	89.96
LVQ	46.10
PNN	97.77
CL	64.31

The system may certainly be extended to other diseases as well. However it is to be noted that the performance of each of the diagnostic system is application dependent, and depends a lot on the kind of data set being used for training and testing. This work can be extended using other algorithm of neural networks. These techniques may be used for other problems like blood pressure identification, heart attack prediction, and fetal health prediction, fetal delivery time structure prediction and blood follow control between pre-natal fetus and maternity abdomen.

4. CONCLUSION

On the basis of the experimental observations of different neural network, HLVQ gives the highest accuracy of 95-98% for diagnosis of breast cancer. Experimental results for test data vectors show that HLVQ has slightly better classification performance and better clustering with efficiency 97.2%, percentage of true positive 95.77%, and percentage of true negative 97.76%. Although speed of the operation is not included here as a measure for comparison.Our thanks to the experts who have contributed towards development of the template.

5. ACKNOWLEDGEMENTS

The authors sincerely acknowledge the Director ABV-IIITM, Gwalior, India for providing facilities to carry out this research work.

6. REFERENCES

[1] World Health Organization International Agency for Research on Cancer (June 2003). "World Cancer Report". http://www.iarc.fr/en/Publications/PDFs-online/World-Cancer-Report/World-Cancer-Report. Retrieved 2009-03-26.

[2] World Health Organization (February 2006). "Fact sheet No. 297: Cancer". http://www.who.int/mediacentre/factsheets/fs297/en/index.html. Retrieved 2009-03-26.

[3] Jemal, A.; Siegel, R.; Ward, E.; Hao, Y.; and "Cancer Statistics,". CA Cancer Journal for Clinicians, 59, 225-249, 2009.

[4] Strax, P., "Make Sure that You Do Not Have Breast Cancer," St. Martin's, NY, 1989.

[5] O. L. Mangasarian and W. H. Wolberg: "Cancer diagnosis via linear programming", SIAM News, Volume 23, Number 5, September 1990, pp 1 & 18.

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

[7] Wolberg, W. H., Street, W. N., and Mangasarian, O. L. (1994). Machine learning techniques to diagnose breast cancer from image-processed nuclear features of fine needle aspirates. Cancer Letters, 77:163-171.

[8] "Cancer Research Statistics: May 2009", 2009, Imperial Cancer Research Campaign, U.K.

[9] Shukla, A.; Tiwari, R.; Kaur, P.; "Knowledge Based Approach for Diagnosis of Breast Cancer", Advance Computing Conference, 2009. IACC 2009. IEEE International, Page(s):6 – 12, 6-7 March 2009.

[10] Janghel R.R , Shukla Anupam, Kala Rahul and Tiwari Ritu "Diagnosis of Breast Cancer by Modular Evolutionary Neural Networks at International Journal of Biomedical Engineering and Technology, Inderscience Enterprises Ltd , Vol. x, No. x, xxxx , pp 1-18, 2010.

[11] Janghel R.R , Shukla Anupam, Kala Rahul and Tiwari Ritu, Intelligent Diagnostic System for the diagnosis and prognosis of Breast Cancer using ANN, Journal of Computing, Volume 2, Issue 12 (Accepted).

[12] Janghel R.R , Shukla Anupam, Kala Rahul and Tiwari Ritu,International Journal of Information Systems and Social Change (IJISSC) (Accepted).

[13] Janghel R.R , Shukla anupam,Tiwari Ritu and Kaur Prabhdeep, Diagnosis of Thyroid Disorders using Artificial Neural Networks on 2009 IEEE International Advance Computing Conference (IACC 2009),Patiala, India, 6–7March 2009,pp:2722-2726.

[14] Janghel R.R , Shukla anupam,Tiwari Ritu and Pritesh Tiwari , "Clinical Decision support system for fetal Delivery using Artificial Neural Network"2009 (NISS) IEEE International Conference on New Trends in

Information and Service Science, June 30 – July 2, 2009, Beijing, China,pp:1070-1075.

[15] Janghel R.R , Shukla Anupam and Tiwari Ritu " Decision Support System for Fetal Delivery using Soft Computing Techniques" (2nd ICIS), 2009 International Conference on Interaction Sciences: Information Technology, Culture and Human, IEEE CPS series ,24-26 November, 2009 - Seoul, Korea, pp: 1514-1519.

[16] Janghel R.R, Shukla Anupam and Tiwari Ritu "Intelligent Decision Support System for Breast Cancer" International Conference on Swarm Intelligence (ICSI 2010) Beijing, China, 12-15 June 2010, ICSI 2010, Part II, LNCS 6146, pp. 351–358, 2010. Springer-Verlag Berlin Heidelberg 2010.

[17] Janghel R.R , Shukla Anupam and Tiwari Ritu "Breast Cancer Diagnosis using Artificial Neural Network Models" (3nd ICIS), 2010 3rd International Conference on Information Sciences and Interaction Sciences, IEEE Explore , June 23-25, 2010, Chengdu, China (Accepted).

[18] Janghel R.R , Shukla Anupam, Tiwari Ritu and Rahul Kala Breast Cancer Diagnostic System using Symbiotic Adaptive Neuro-evolution (SANE), The lnternational Conference of Soft Computing and Pattern Recognition(SoCPaR2 010), IEEE Explore, Dec 07-10, 2010 ,France.