Volume 2 No.7, July 2012 ISSN 2224-3577
International Journal of Science and Technology

http://www.ejournalofsciences.org

Genetic-Neuro Approach for Disease Classification

P.Venkatesan [1], V. Premalatha [2]
[1] National Institute for Research in Tuberculosis, Indian Council of Medical Research, Chetpet, Chennai-31.
[2] Department of Mathematics, St. Joseph's College of Engineering, Chennai-119.

ABSTRACT

In this study, we introduce classification using Artificial Neural network whose architecture is trained by Back propagation algorithm and input attributes are selected by evolutionary based Genetic algorithm. Feature selection plays vital role in the applications of Machine learning algorithms. For feature selection many approaches are used and the evolutionary based algorithms prove to be one of the efficient methods. We have used a breast cancer dataset for the evaluation of the new approach. The experimental result shows that the genetic-neuro system classification performs better than the conventional neural network.

Keywords: *Classification, Feature selection, Genetic algorithm, Neural Network*

1. INTRODUCTION

The complexity and the volume of data are strictly increasing in many disciplines particularly in biomedical sciences. To explore and identify patterns in the huge volume of data, the machine learning based approaches play a vital role. Machine learning is a subfield of artificial intelligence and its goal is to program computers to solve any problem with the help of past experience [1], [2]. It is also of designing and developing algorithms for analyses of data for classification, prediction and association. In machine learning, the performance of the classification depends upon the input attributes. Neural network is used for classification of diseases based on features of the patients [3]. In Multi-Layer Perceptron's, number of layers and the total number of neurons present in each layer is estimated by the input attributes. So, selecting input attributes which contribute significantly to the classification problems plays vital role in the above tasks. In this paper the evolutionary based Genetic algorithm is used for best feature selection to provide as inputs to neural network.

Breast cancer is one of the deadly disease, most common in women [4],[5]. Universally it caused 458,503 deaths in women in 2008[6]. In Western countries, the survival rates are very high when compared to developing countries. Although there are tests involved to diagonize the early stage of breast cancer [7], the experts go for computer-aided diagnosis (CAD) for confirming their prediction. This CAD helps to improve their prediction efficiency and accuracy [8-20]. In breast cancer, the prime target is to differentiate between malignant and benign case. CAD should be trained well with the past experience to predict whether the case is malignant or benign. It should also be user-friendly, so that the experts can have the classification with explanation.

2. NEURAL NETWORKS

Neural network is an extremely simplified version of the brain and nervous system as shown in figure 1. Artificial neural network is composed of many neurons that will join together to serve as an information processing unit. It mainly incorporates the two fundamental components of biological neural nets-nodes and weights. ANN is used for pattern recognition or data classification using learning process, by adjusting synaptic connection between neurons. Here we are going to discuss Multi Layer Perception, which is a type of network as shown in figure 2 and also the most widely used neural network architecture.

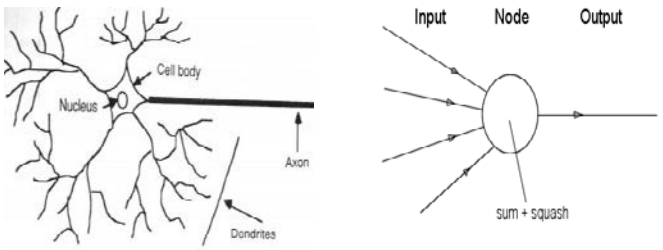

Fig 1: Human brain Vs Neural Net

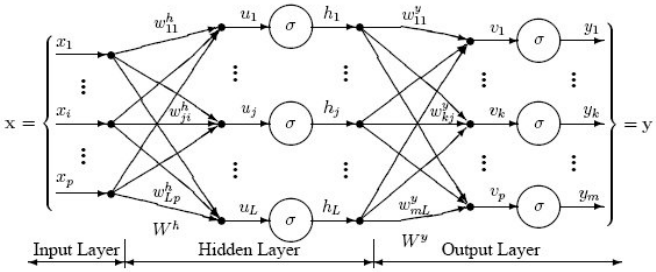

Fig 2: Architecture of MLP

A multilayer perceptron is a general feed-forward artificial neural network, which solves non-separable problems.

It consists of more than one hidden layer of neurons and the updates in the layers are exhibited starting from the input to the output [21]. A binary sigmoid function defined by $\frac{1}{1+e^{-x}}$ is the activation function considered for each node. For the incoming signals, each neuron will find a weighted sum to get the net input. Using this input, the sigmoidal activation function gets the required activation value to the corresponding neuron. The aim of NN training is to minimize the network error, which is calculated using a mathematical function. The value of the error function is used to rate the quality of the neural network. For a Multi Layer Perceptron the error function [22] is given by:

$$E = \sum_{i=1}^{P} \underbrace{(d_i - f(x_i, w))^2}_{Closeness\ to\ data}. \quad \textbf{(1)}$$

Neurons have weighted inputs, threshold values, activation and an output as shown in Figure 3.

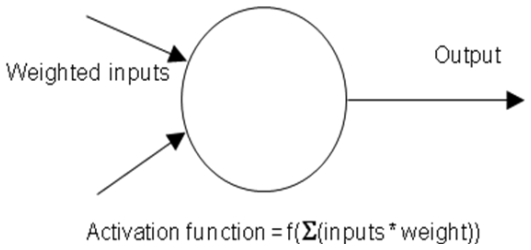

Fig 3: Neuron model

2.1 Back Propagation

Back propagation is a supervised learning method, most commonly used for training ANN. It first modifies the output neuron's synaptic weight and then start modifying the synaptic weights of all other layer starting from p-1 to the first layer and propagates backward the local error terms. Consider P vector pairs (x_1,y_1), (x_2,y_2),… (x_P,y_P) which belongs to the functional mapping $y=\phi(x)$ such that $x\in R^N$, $y\in R^M$. The network will learn an approximation $o=y'=\phi'(x)$ after the training. The single training vector pair's relevant equation are given in order [23],[24]. Let the input vector sent through the input layer be $x_p = \left(x_{p_1}, x_{p_2}, ..., x_{p_N}\right)^t$, whose values are also distributed to the hidden layers. Let the j[th] hidden units net input be:

$$net^h_{p_j} = \sum_{i=1}^{N} w^h_{ji} x_{p_i} + \theta^h_j \quad (2)$$

where w^h_{ji} is the weight carried from the connection from i[th] node to j[th] node, θ^h_j denotes the bias term and h denotes the hidden layer's quantities. If the activation of the node and the net input are equal then

$$i_{pj} = f^h_j(net^h_{p_j}) \quad (3)$$

The output nodes equations are

$$net^o_{pk} = \sum_{j=1}^{L} w^o_{kj} i_{pj} + \theta^o_k \quad (4)$$

$$o_{pk} = f^o_k(net^o_{pk}) \quad (5)$$

output layers quantities is denoted by superscript "o".
The error terms for the output units is given by

$$\delta^o_{pk} = (y_{pk} - o_{pk}) f^{o'}_k(net^o_{pk}) \quad (6)$$

The error terms for the hidden layers, calculated before the connections weight to the output-layer units updated is given by

$$\delta^h_{pj} = f^{h'}_j(net^h_{pj}) \sum_k \delta^o_{pk} w^o_{kj} \quad (7)$$

Update weights on the output layer is given by

$$w^o_{kj}(t+1) = w^o_{kj}(t) + \eta \delta^o_{pk} i_{pj} \quad (8)$$

where η is the learning-rate parameter.

Update weights on the hidden layer is given by:

$$w^h_{ji}(t+1) = w^h_{ji}(t) + \eta \delta^h_{pj} x_i \quad (9)$$

But the error term is a measure of networks learning capacity and if it is considerably small for each training vector pairs then the training will be stopped. The error term is given by:

$$E_p = \frac{1}{2} \sum_{k=1}^{M} \delta^2_{pk} \quad (10)$$

3. GENETIC ALGORITHM

Genetic algorithms are adaptive randomized search, optimization technique based on the evolutionary ideas of natural selection with stochastic operators. John Holland's pioneering work, shows how the highly parallel technique,

genetic algorithm is applied to solve wide variety of problems [25]. Their robust performance gives hope for many optimization problems, design, control and machine learning applications [26]. In optimization problems, how genetic algorithm was successfully applied was described in [27],[28],[29].

Genetic algorithms are a part of evolutionary computing, which is a subfield of artificial intelligence. Biologically speaking, every organism has its own rule to describe how it is built up from the tiny building blocks. These rules are encoded in the genes of every organism. Such genes are connected to form a long strings said to be Chromosomes. Highly fit chromosomes are selected for reproduction to produce offspring. The offspring's may undergo mutation with less chance. This is the way in which, lives on earth has evolved.

Genetic algorithm deals with pool of solution (i.e., population of strings) called the chromosomes. Genetic algorithm operates on encoded representation of the candidate solutions to an optimization problem, which is equivalent to those chromosomes of the individuals in nature. Most probably solutions are encoded in binary strings of 0s and 1s, also other encodings are possible. For evaluating each chromosome a fitness function should be assigned then the fittest chromosomes are selected to undergo the operations crossover and mutation to produce a new pool of solutions. The process will be continued till we get more hightly fittest chromosomes in the pool of solutions. Other stopping criteria are number of generations, almost stable value of the best fitting individual and almost stable value of the average fitness of the population. To know more about basic principles of GA, the principles and applications are discussed by many authors [30]-[35].

Genetic parameters are population size, probability of crossover and probability of mutation. Population size plays important role in affecting efficiency of the algorithm. It should neither be small in size nor large. If it is small then it will yield to poor performance. At the same time very large population size will make the convergence rate to be slow. If the probability of crossover is high then we get the offspring's quickly. At the same time the crossover rate is not suppose to be too high or low. Mutation rate is always low, which will ensure that no bit position is stuck to single values.

3.1 Schema and Schema Theorem

If the similarities among chromosomes are allowed to explore by a template, then it is said to be schema. In the gene alphabet a "don't care" symbol is introduced. into the gene alphabet, to represent all the strings which differ only in '*' position. The schema (* 1 1 1 1 0 0 1 0 0 1) matches the two strings (0 1 1 1 1 0 0 1 0 0 1), (1 1 1 1 1 0 0 1 0 0 1). If m(S,t) denote the expected number of individuals belonging to schema S at time t then

$$E[m(S,t+1)] \geq m(S,t)\frac{f(S,t)}{\bar{f}(t)}\left[1-p_c\frac{\delta(S)}{l-1}-p_m o(S)\right] \qquad (11)$$

Where $\bar{f}(t)$ is the average fitness of the population at time t, p_c is the probability of crossover and p_m is the probability of mutation. The schema theorem states that short, low-order, above average schemata receives exponentially increasing trials in subsequent generations of a Genetic algorithm. Such schemata are called building blocks. Such building blocks together achieve the high performance of genetic algorithm.

4. RESULTS AND DISCUSSION

The breast cancer data set was obtained from UCI Machine Learning Repository [36], [37]. The input variables for the classification of breast cancer disease are given in Table 1.

Table 1: Variables of Breast Cancer Disease

Sl.No.	Input variables
1.	Sample code number
2.	Clump Thickness
3.	Uniformity of Cell Size
4.	Uniformity of Cell Shape
5.	Marginal Adhesion
6.	Single Epithelial Cell Size
7.	Bare Nuclei
8.	Bland Chromatin
9.	Normal Nucleoli
10.	Mitoses

Neuro-Intelligence tool [38] is used to model the breast cancer data. The data consists of 659 cases each consisting of the variables given in Table 1 and the last column consists of the target variable (Malignant or Benign). The dataset is divided into training set, validation set and test set using random data partition method given in Table 2.

Table 2: Data Partition Set

S.No	Data Partition set	Records	Percentage
1.	Training set	449	68.13%
2.	Validation set	105	15.93%
3.	Test set	105	15.93%
	Total	659	100%

4.1 Prediction Using NN and GA-NN

In feature selection using Genetic algorithm the parameters taken into account are given in Table 3.

Table 3: Parameters used in GA

Search Method	Genetic Algorithm
Smoothing	0.01
Unit Penalty	0.001
Sampling	100
Population	50
Number of generation	50
Probability of crossover	0.9
Probability of mutation	0.1

Using Genetic algorithm, the best features combination found are Clump Thickness, Uniformity of Cell Shape, Marginal Adhesion, Bare Nuclei and Normal Nucleoli. The prediction accuracy using NN and GA-NN are given in Table 4. The figure 4 gives errors with respect to training set, validation set and the best network. The figure 5 gives the ROC for testing.

Table 4: Prediction accuracy using NN and GA-NN

	Neural Network		**GA-NN**	
	Training	Validation	Training	Validation
CCR%	93.986637	96.190476	97.327394	98.095238
Network error	0.628534	-	0.069682	-
Architecture	10-5-1		5-2-1	

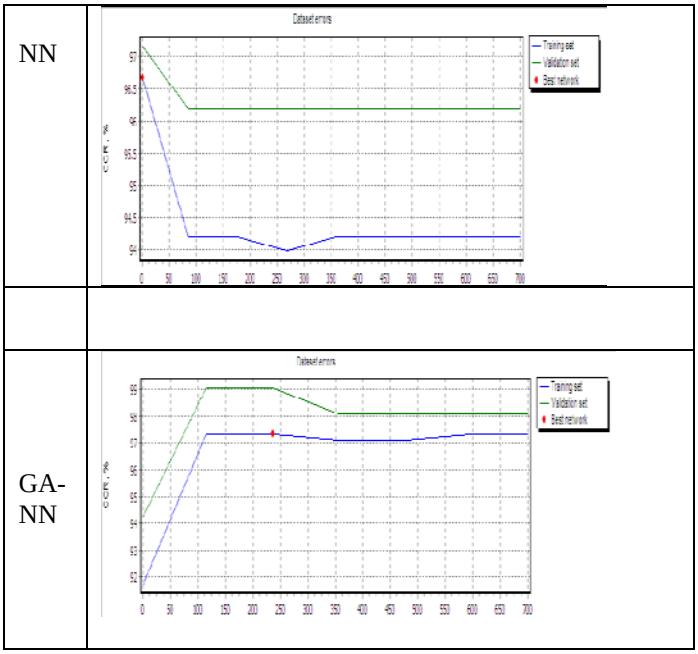

Figure 4: Errors under Training, Testing and Validation

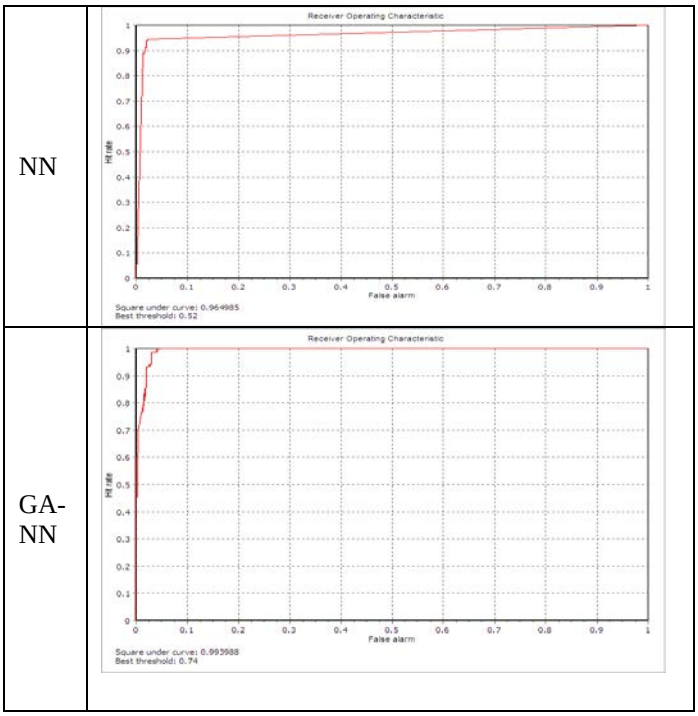

Figure 5: ROC for testing

5. CONCLUSION

In this paper genetic-neuro approach is applied for classification of breast cancer. The experimental results show that the combination of best features selected using GA to the input of ANN improves the performance of ANN. Also the reduced input features results in the reduction of size of the neural network and the complexity of neural network. The Wisconsin breast cancer classification problem was studied by many authors using fuzzy-genetic approach [39], Support Vector Machines combined with feature selection [40] and our genetic-neuro approach performance is similar to them and agrees with their classification. Further studies are needed to confirm these findings.

REFERENCES

[1] Ethem Alpaydin(2006), Introduction to Machine Learning, Prentice – Hall of India.

[2] T. Mitchell (1997). Machine Learning, McGraw Hill. ISBN 0-07-042807-7.

[3] D.B.Fogel, E.C.Wasson, and E.M.Boughton(1995). Evolving Neural Networks for Detecting Breast Cancer, Cancer Letters, Vol. 96,pp. 49-53.

[4] C.C.Boring , T.S.Squires , T.Tong . Cancer statistics 1993. CA: Cancer J. Clinicians 1993; 43:7-26.

[5] P.A.Wingo, T.Tong , S.Bolden . Cancer statistics 1995. CA: Cancer J. Clinicians 1995; 4523-30.

[6] "World Cancer Report". (2008):International Agency for Research on Cancer. Retrieved 2011-02-26.

[7] S.A.Feig (1988). Decreased breast cancer mortality through mammographic screening: Results of clinical trials.Radiology 167:659-665

[8] C.J.Vyborny, M.L.Giger(1994). Computer vision and artificial intelligence in mammography. AJR 162:699-708.

[9] H.P.Chan, K.Doi, C.J.Vyborny , et al.,(1990). Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. Invest Radiol 25: 1102-1 110.

[10] N.Karssemeijer, J.H.Hendriks(1997). Computer-assisted reading of mammograms. Eur Radiol 7:743-748.

[11] W.P.Kegelmeyer, J.M.Pruneda, P.D.Bourland, A.Hillis, M.W.Riggs, *M.L.*Nipper(1994). Computer-aided mammographic screening for spiculated lesions. Radiology 191:331- 337.

[12] H. Kobatake and Y. Yoshinaga.(1996). Detection of spicules on mammogram based on skeleton analysis, IEEE Trans. Med. Imaging 15, 234–245.

[13] C. H. Chen and G. G. Lee(1997): On digital mammogram segmentation and micro calcification detection using multiresolution wavelet analysis, Graph. Models Image Process 59, 349–364.

[14] L. M. Bruce and R. R. Adhami(1999): Classifying mammographic mass shapes using the wavelet transform modulus-maxima method, IEEE Trans. Med. Imaging 18, 1170–1177.

[15] H.P. Chan *et al.*,(1999): Improvement of radiologists' characterization of mammographic masses by using computer-aided diagnosis: An ROC study, Radiology 212, 817–827.

[16] Z. Huo, M. L. Giger, and C. E. Metz.,(1999) : Effect of dominant features on neural network performance in the classification of mammographic lesions, Phys. Med. Biol. 44, 2579–2595.

[17] Z. Huo, M. L. Giger, C. J. Vyborny, D. E. Wolverton, and C. E. Metz., (2000): Computerized classification of benign and malignant masses on digitized mammograms: A study of robustness, Acad. Radiol. **7**, 1077– 1084.

[18] N. R. Mudigonda, R. M. Rangayyan, and J. E. L. Desautels.,(2001): Detection of breast masses in mammograms by density slicing and texture flow-field analysis, IEEE Trans. Med. Imaging 20, 1215–1227.

[19] B. Sahiner *et al.*, (2001): Computer-aided characterization of mammographic masses: Accuracy of mass segmentation and its effects on characterization, IEEE Trans. Med. Imaging 20, 1275–1284.

[20] S.-C. B. Lo, H. Li, Y. Wang, L. Kinnard, and M. T. Freedman., (2002): A multiple circular path convolution neural network system for detection of mammographic masses, IEEE Trans. Med. Imaging 21, 150–158.

[21] Gary William Flake.,(2000).The Computational Beauty of Nature-Computer Explorations of Fractals, Chaos, Complex Systems, and

Adaptation, MIT Press.

[22] Vojislav Kecman(2001). Learning and Soft Computing- Support Vector Machines, Neural Networks, and Fuzzy Logic Models, MIT Press.

[23] D. E. Rumelhart, G. E. Hinton, and R. J. Williams(1986). Learning internal representation by error propagation, in Parallel Distributed Processing: Explorations in the Microstructures of Cognition. Cambridge, MA: MIT Press, vol. 1

[24] James A. Freeman, David M. Skapura (1991). Neural Networks algorithms, applications and programming technique, Addison-Wesley Publishing, Inc.

[25] J.Holland(1975): Adaptation in Natural and Artificial Systems. The University of Michigan.

[26] U. Maulik and S.Bandyopadhyay (2000). Genetic algorithm-based clustering technique. Pattern Recognition 33, 1455-1465.

[27] E.Aarts and J.K.Lenstra (Eds.) (1997) Local Search in Combinatorial Optimization, John Wiley & Sons, Chichester.

[28] C.R.Reeves (Ed.) (1993). Modern Heuristic Techniques for Combinatorial Problems, Blackwell Scientific Publications, Oxford, UK; re-issued by McGraw-Hill, London, UK (1995).

[29] C.R.Reeves (1996). Heuristic search methods: A review. In D.Johnson and F.O'Brien (1996) Operational Research: Keynote Papers 1996, Operational Research Society, Birmingham, UK, pages 122-149. ISBN 0903440164.

[30] D.Beasley, D.R. Bull, and R.R. Martin(1993). An overview of genetic algorithms: Part 1-Fundamentals, Univ. comput., vol. 15, no. 2, pp.58-69.

[31] D.E. Goldberg (1989). Genetic algorithms in Search, Optimization and Machine Learning. Reading,MA: Addison-Wesley.

[32] Z. Michalewicz(1994). Genetic algorithms + Data Structures = Evolution program, 2nd Ed. Berlin: Springer-verlag.

[33] Michael. D. Vose(1999). The Simple Genetic Algorithm-Foundations and theory. MIT Press.

[34] L.Davis (1991). Handbook of genetic algorithm. San Nostrand Reinhold, New York.

[35] M. Mitchell (1997). An introduction to genetic algorithms, the MIT Press, Cambridge, Massachusetts.

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

[37] William H. Wolberg and O.L. Mangasarian (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology, Proceedings of the National Academy of Sciences, U.S.A., Volume 87, pp 9193-9196.

[38] Neuro-Intelligence using Alyuda. Source available at www.alyuda.com. Last accessed 10 May 2008.

[39] C. A. Pena-Reyes, M. Sipper (1999). A fuzzy-genetic approach to breast cancer Diagnosis, Carlos Andres Pen a-Reyes, Moshe Sipper, Artificial Intelligence in Medicine, vol. 17, pp. 131–155.

[40] M. F. Akay (2009). Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications, vol. 36, Issue 2, part 2, pp. 3240–3247.